UNIVERSIDAD LA SALLE

FACULTAD DE NEGOCIOS

Con Reconocimiento de Validez Oficial de Estudios de la Secretaría de Educación Pública, según Decreto Presidencial de fecha 29 de mayo de 1987

TESIS

PREDICCIÓN DE VARIACIÓN DE PRECIO DE UNA ACCIÓN EN EL MERCADO BURSÁTIL MEXICANO MEDIANTE MODELOS DE REGRESIÓN

QUE PARA OBTENER EL TÍTULO DE

LICENCIADO EN ACTUARÍA

PRESENTA:

LUIS ANTONIO SÁNCHEZ ARRIAGA

Asesora Dra. María del Carmen Lozano Arizmendi

Ciudad de México, México

Octubre de 2020

A mis padres por su cariño y apoyo incondicional.

A mi abuelo Don Enrique Sánchez por tanto.

Agradecimientos

El presente trabajo no hubiera sido posible sin la guía y asesoramiento que me brindó la Dra. María del Carmen Lozano Arizmendi, tomando este proyecto como asesor de tesis.

Un especial agradecimiento al profesor Mtro. Adolfo Rangel Díaz de la Vega por haberme instruido en las Finanzas Corporativas y en el Análisis Bursátil que en consecuencia resultaron en la investigación aquí presente.

Mi agradecimiento y admiración a toda la plantilla de profesores de la Lic. en Actuaría de la Universidad La Salle México por su arduo trabajo en elevar el nivel de la carrera y que en su ejercicio me han ilustrado de una manera enorme.

Ciudad de México a 16 de abril de 2021

MTRA. ANA MARCELA CASTELLANOS GUZMÁN
DIRECTORA DE GESTIÓN ESCOLAR
UNIVERSIDAD LA SALLE
P R E S E N T E

Le informo que el (la) C.

LUIS ANTONIO SÁNCHEZ ARRIAGA

Egresada(o) de la Facultad de Negocios
de la **UNIVERSIDAD LA SALLE**, de la Licenciatura en:

ACTUARÍA

Con reconocimiento de validez oficial de estudios de la Secretaria de Educación Pública Según Decreto Presidencial de fecha 29 de mayo de 1987.

Ha elaborado la tesis titulada: "**PREDICCIÓN DE VARIACIÓN DE PRECIO DE UNA ACCIÓN EN EL MERCADO BURSÁTIL MEXICANO MEDIANTE MODELOS DE REGRESIÓN.**"

De conformidad con la modalidad para la obtención de título aprobada para esta Licenciatura de acuerdo a lo establecido en el Reglamento General de las Universidades La Salle Integrantes del Sistema Educativo de las Universidades la Salle.

Cumplió con todos los requisitos y el trabajo que fue elaborado bajo mi conducción, tiene la calidad suficiente para ser la base de sustentación de su Examen Profesional por lo que se le autoriza presentarlo.

Mtro. José Ramón Barreiro Iglesias
Director Facultad de Negocios

Índice general

Capítulo 1

Introducción

1.1. Sobre la predicción del precio de una acción

Diversos autores y expertos en el mercado bursátil aseguran que el comportamiento del precio de una acción es completamente caótico y es casi imposible predecir el precio futuro de una acción (Bodie, Kane y Marcus, 2011). Por lo que se han desarrollado diferentes algoritmos y métodos con el fin de estimar el precio futuro de una acción bursátil. Entre los métodos más comúnes están, por ejemplo, la fórmula de Black-Scholes, y herramientas basadas en inteligencia artificial, que utilizan métodos basados en el reconocimiento de patrones para estimar si una acción va a bajar o a subir de precio.

Desde el punto de vista del análisis fundamental, el precio de una acción está fuertemente relacionado con el desempeño de la compañía. Esto es, el inversionista obtiene información de los estados financieros de la compañía. En este sentido, el inversionista puede tomar como parámetro el valor intrínseco de una acción (cociente entre capital y número de acciones emitidas) para estimar el precio de una acción. En otras palabras, según el análisis fundamental, un parámetro a tomar en cuenta es el precio teórico de una acción ya que si éste supera al valor de mercado significa que el mercado está subestimando dicha acción y por lo tanto es buena oportunidad de compra (se está comprando una acción a un precio menor del real y la expectativa es que la acción eventualmente tome el valor teórico, y entonces se obtenga el diferencial de ganancia). De lo contrario, si el precio teórico de una acción está por debajo del precio de mercado, según el análisis fundamental, es buena oportunidad de venta (se está vendiendo un título por más de lo que realmente vale, por lo tanto, eventualmente, el mercado corrige la tendencia alcista y se llega al precio teórico, en otras palabras, se vende el título antes de que la acción baje de precio y se obtiene dicho diferencial).

La dificultad de analizar el comportamiento del mercado bursátil recae en la gran cantidad de variables que están en juego. Considere el siguiente escenario para ilustrar el argumento anterior: suponga que una empresa X que cotiza en la Bolsa Mexicana de Valores (BMV) ha tenido dificultades al obtener ganancias, y en consecuencia, ha mostrado un pobre desempeño financiero. Sin embargo, han observado una oportunidad de inversión en un producto que robará la atención del mercado y dejará rendimientos atractivos para los accionistas. Los accionistas más informados, no dudarán en tomar ventaja y comprar acciones de X, antes de que dicho producto incremente el interés del mercado de poseer una porción de X. Debido al pobre desempeño de X, el valor teórico de sus acciones estará muy por debajo al precio en BMV. De este ejemplo se puede inferir que no existe precio justo para un título bursátil, el único precio *just* será el que el mercado esté dispuesto a pagar. Desde luego, el tenedor de la acción planteado previamente no tiene un conocimiento acerca del precio futuro de la acción, así que opta por especular y retener su acción hasta observar un precio de mercado que le proporcione una ganancia.

Es importante que se haga hincapié en que, si una acción va a subir o no de precio, el único parámetro a seguir es el movimiento del mercado. Desde luego existen factores que un inversionista puede seguir para tener un bosquejo del comportamiento de la acción. Para la BMV, se puede contar con el índice de Precios y Cotizaciones (IPyC), el cual es un índice compuesto por la elaboración de un portafolio de acciones conformado por las empresas más representativas del mercado. En resumen, el IPyC mide la variacón porcentual de la suma de del valor de capitalización de cada serie accionaria parte de la muestra, de un día a otro (Ladrón de Guevara, 2004).

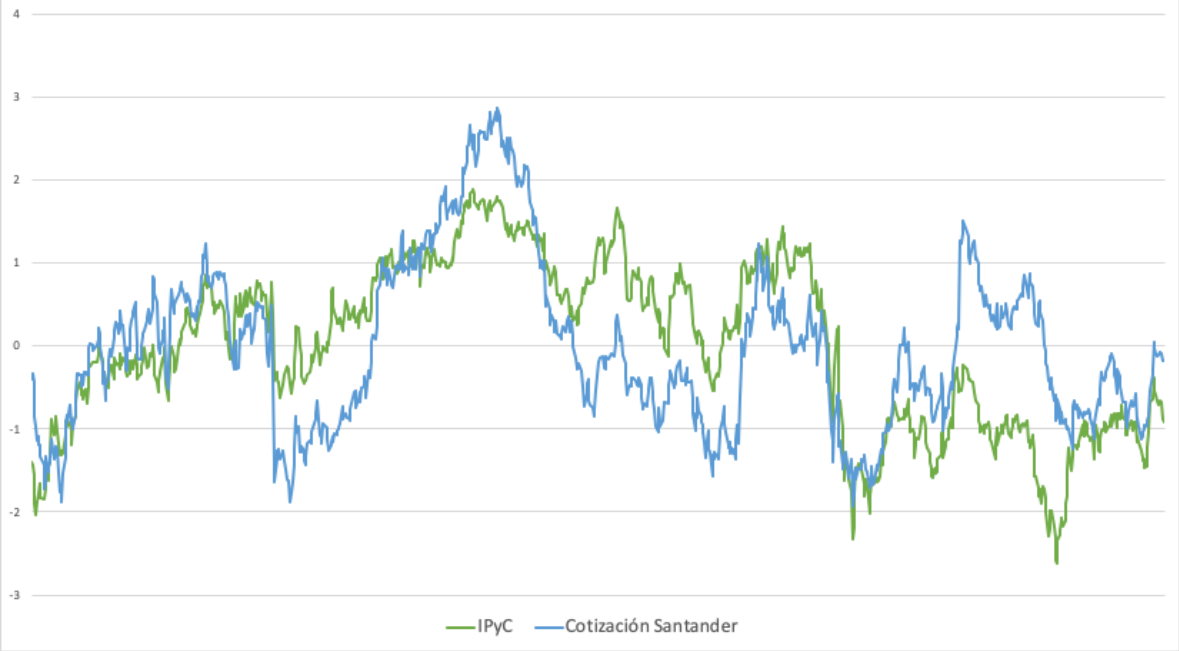

Figura 1.1: Comparación de Santander contra el mercado en el periodo enero 2016 a enero 2020. Fuente: Elaboración propia con datos normalizados de Yahoo Finance.

En la Figura 1.1 se muestra la trayectoria de precios durante un año (978 días hábiles) de la cotización de cierre de Santander y el IPyC (ambos normalizados) desde el 4 de enero de 2016 hasta el 1 de enero de 2020. Es notorio que los movimientos de los precios de la acción siguen una trayectoria muy similar a la del IPyC. De ahí que un parámetro a seguir para cualquier acción, es el movimiento en general del mercado, además de otras variables como empresas del mismo sector, un portafolio conformado por todas las empresas del sector, tipo de cambio, tasa de fondeo gubernamental, entre otros.

Sin embargo, tal indicador proporciona un panorama general del comportamiento del mercado, sin tomar en cuenta que los diversos sectores de las compañías incluidas en el IPyC no se comportan de la misma manera. Por ejemplo, si un inversionista busca invertir en una acción de Santander, consultar los movimientos recientes del sector financiero es una práctica razonable. En contraste, el inversionista podría consultar los movimientos del sector energético, el cual tendrá poca o nula relación con el precio de la acción deseada. Desde este punto de vista, hay tres factores a tomar en cuenta:

I. El desempeño propio de la emisora durante un periodo de tiempo determinado (diariamente, quincenalmente, etc.).

II. El movimiento del mercado en general (IPyC).

III. El comportamiento del sector al que pertenece la emisora.

1.2. Planteamiento del problema

Fenómenos financieros como la bancarrota han sido estudiados por décadas desde diferentes enfoques y se han desarrollado gran variedad de métodos para estudiar estos fenómenos. A decir, se utilizaron métodos estadísticos basados en información financiera de empresas para predecir la bancarrota, por ejemplo, (Altman, 1968) y (Deakin, 1972). No obstante, estos métodos tienen limitaciones importantes con relación a los supuestos estadísticos que deben cumplir los datos.

El precio de una acción es un fenómeno financiero para el que también se han propuesto diversas herramientas y métodos enfocados a determinar el precio futuro de una acción. Por un lado, se tienen los métodos clásicos como el modelo de valoración de activos financieros (CAPM por sus siglas en inglés) o la teoría de valuación de precios por arbitraje (APT por sus siglas en inglés). Ambos utilizan métodos clásicos de valuación de activos: a través de conjuntos eficientes y análisis de sensibilidad, calculan el rendimiento esperado de un activo (en este caso una acción). Posteriormente, se utiliza dicho rendimiento para valuar los flujos futuros asociados a la acción (Ross, Westerfeld y Jordan, 2010).

En consecuencia, se espera que un método para los precios de acciones del sector financiero permita tomar mejores decisiones de inversión para que cualquier inversionista o interesado en general sea capaz de visualizar un escenario que maximice su ganancia.

En pocas palabras, el problema de predecir el precio de una acción o hacer inferencias sobre el futuro de una acción conlleva algunas dificultades, entre ellas, la de analizar un gran volumen de información y la de determinar la desviación y el error asociado a las predicciones una vez que el modelo ha arrojado algún resultado y con ello tener los argumentos suficientes para tomar una decisión financiera (compra o venta).

Tomando en cuenta lo anterior, la presente investigación hace un esfuerzo por al menos proponer una solución ante estos dos problemas. Por un lado, se introducen modelos de aprendizaje de máquina (particularmente regresión lineal y regresión logística) que con base en datos históricos, permiten ajustar parámetros para que el modelo sea replicado en el futuro. Por otro lado, en la presente investigación se definen y se ponen en práctica diversas métricas de error y validación de las predicciones del modelo.

El conocimiento del impacto y utilidad de modelos basados en regresión aplicados en las finanzas es de suma utilidad para obtener predicciones buenas con base en factores internos del mismo mercado. La propuesta de nuevos métodos para predicciones en el mercado financiero podrá reducir la incertidumbre y acercarse a la realidad financiera.

1.3. Objetivo general

Esta investigación tiene como objetivo estimar la probabilidad de alza de una acción, condicionada a los movimientos de precios de las distintas empresas que cotizan en el mismo sector, utilizando regresión logística. Esta probabilidad fungirá como parámetro para clasificar si una acción subirá de precio o no. De tal manera que un inversionista o interesado en general sea capaz de visualizar un escenario que maximice su ganancia. Dada la gran cantidad de información disponible (histórica y actual), se presentará una alternativa solución a un problema cuya dificultad recae en el análisis multifactorial que requiere; con ello se brindará al público inversionista y académico una alternativa para estimar la conveniencia de compra o venta de una acción, dado el movimiento del mercado.

Esta investigación propone la regresión logística como un modelo para estimar la probabilidad de alza o baja del precio de acciones de Santander. Se eligió regresión logística puesto que predice la probabilidad de un resultado que solo puede tener dos valores (dicotómico), en este caso alza o baja del precio de una acción. La predicción se basa en el uso de uno o varios predictores (numéricos y categóricos).

La motivación principal de utilizar la regresión logística como modelo de clasificación y predicción es su naturaleza probabilística, pues se estima la función de distribución conjunta de manera empírica de que una acción en particular suba de precio, dado los movimientos del mercado, además la practicidad de implantación a través de diversos paquetes estadísticos.

Por otro lado, la naturaleza dicotómica de la variable de respuesta contribuye a que las métricas de error y validación sean interpretables y fáciles de entender.

Todos los cálculos, gráficas y validaciones se llevan a cabo a través del lenguaje *R* que es comúnmente usado en temas de aprendizaje de máquina y estadística computacional. Particularmente, se hace un uso extenso de la librerías *tidyverse, glm, roc.* El paquete *tidyverse* se usa para manipulación de datos y gráfcas, *glm* para los modelos de regresión y finalmente *roc* para las métricas de validación y precisión. Además, en *Python* se utilizaron los paquetes *pandas*, $yfinance$, *plotly* y *scikitlearn*.

1.4. Objetivos específicos

I. **Revisión de literatura**: Definir conceptos financieros, así como el contexto del mercado de valores mexicano y mostrar aplicaciones previas de algoritmos en el mercado bursátil. Posteriormente, construir el estado del arte en torno a la predicción de variación de precios de acciones en el mercado de valores utilizando técnicas de aprendizaje de máquina.

II. **Definir el concepto de regresión logística y estimación de parámetros**: Definir los conceptos y teorías relacionadas con la regresión logística, así como estimación de parámetros e interpretación. Esto permitirá proponer un modelo que permita estudiar la variabilidad hacia el alza o baja de una acción condicionada a los movimientos de empresas del mismo sector. La regresión logística es particularmente compatible con el objetivo de la investigación, puesto que los resultados están en términos de probabilidad condicional.

III. **Presentar análisis sobre la emisora Santander México (BSMX)**: Puesto que, para ilustrar la investigación, se ha escogido el sector financiero de BMV, será importante ahondar en la situación financiera y contexto de Santander, así como de las empresas del mismo sector.

IV. **Mostrar análisis gráfico**: Un proyecto de aprendizaje de máquina es mucho más enriquecedor cuando es acompañado de un profundo análisis gráfico. Por tal razón, a lo largo de los capítulos se presentarán diferentes visualizaciones para un mejor entendimiento de los datos y del problema.

V. **Discriminar variables**: Con la base de datos creada, hacer un análisis de regresión lineal múltiple explicando el precio de la empresa de interés a través de los precios de otras emisoras del mismo sector. Esto con la meta de discriminar variables que tengan poca relación lineal en el precio de la emisora de interés.

VI. **Clasificar observaciones**: Considerando las variables dictadas por el modelo de regresión, aplicar un modelo de regresión logística, explicando el alza o baja de la acción en cuestión, dada la variación del mercado. Puesto que las respuestas del modelo vienen dadas por probabilidades, se establecerá un umbral en el cuál se tome a la respuesta como positiva.

VII. **Validar predicciones**: Una vez ajustado el modelo de regresión logística, se utilizará otra conjunto de datos, con el objetivo de medir la calidad y precisión de las predicciones.

VIII. **Interpretar el modelo**: Validados los resultados, interpretar y concluir, acorde al mercado de valores mexicano.

1.5. Justificación del tema

Como se ha señalado anteriormente, el precio de las acciones se rige por las leyes de oferta-demanda del mercado, por lo tanto, los precios reaccionan positiva o negativamente ante factores internos del mercado, por ejemplo, condiciones propiamente del país en el que se ofertan (inflación, tipo de cambio, calificaciones, etc.). Así como a factores propios de las emisoras, como sus precios de cotización diaria, desempeños financieros, etc.

Recolectando precios de cotización diarios de emisoras del mismo sector e índice de precios y cotizaciones (IPyC), esta investigación se centrará en predecir la probabilidad de alza o baja de una acción, dado el movimiento de tales factores del mercado. Tal movimiento vendrá dado por los rendimientos diarios de los indicadores financieros mencionados.

Se pretende utilizar técnicas de aprendizaje supervisado tratando de predecir la respuesta positiva o negativa de una emisora del sector financiero (0 si sube, 1 si baja), dados los rendimientos de acciones de emisoras del mismo sector, así como IPyC, tipo de cambio y tasa libre de riesgo. Una vez hechas las predicciones, se medirá la precisión del modelo con matrices de confusión, así como la curva ROC (Receiver Operating Characteristic) y su respectivo AUC (Area Under The Curve), que son métricas necesarias para presentar los resultados de un modelo de aprendizaje (Marsland, 2015).

Realizar pronósticos sobre activos financieros es de gran utilidad para cualquier institución privada o gubernamental que esté involucrada en el mercado directa o indirectamente. En el mercado existen diversos factores que influyen en el desempeño de cualquier economía. Entre ellas está la tasa libre de riesgo, la inflación y el producto interno bruto. El mercado de valores es de suma importancia, pues las principales compañías cotizan en él.

Como se mencionó anteriormente, en un mercado como el mexicano los precios de cotización no son fijos pues se rigen por oferta y demanda. De tal manera que *a priori*, invertir en el mercado de valores es básicamente una apuesta ya que no hay manera de saber si una acción subirá o bajará de precio (sin acudir a la especulación).

En consecuencia, la contribución principal de esta investigación es proponer una técnica para evaluar la conveniencia de invertir en una determinada acción. Se diseñará una serie de pasos para evaluar dicha conveniencia de manera objetiva con la finalidad de evitar apelar a la especulación para invertir o no en la bolsa. Por otro lado, se exhibe cómo el aprendizaje de máquina puede ser de gran apoyo para cualquier institución o persona para medir el desempeño de la cotización de una acción.

1.6. Hipótesis de Investigación

La aplicación de los modelos de regresión en el estudio de las acciones del mercado de valores en México parten de dos supuestos muy importantes y que fueron desarrollados a lo largo de la investigación. Éstos pueden ser formulados en las siguientes hipótesis de trabajo:

I. Los precios de las acciones en el mercado de valores mexicano, responden a factores internos del mercado, particularmente al cambio de precios en acciones del mismo sector.

II. El modelo de regresión logística puede ser aplicado para evaluar la conveniencia de compra/venta de acciones en el corto plazo.

1.7. Estructura

En el Capítulo 2 de esta tesis, se presenta una revisión de la literatura. Se define mercado financiero así como los conceptos básicos relacionados con este concepto. Se aborda el mercado accionario tema central de esta investigación. También se aborda la entidad bancaria Santander que será el objeto

de estudio de esta investigación. Posteriormente, en el Capítulo 3 se establecen los modelos clásicos y modernos que existen para el estudio de los mercados financieros. El diseño de la investigación, variables utilizadas y el tamaño de la muestra se discute en el Capítulo 4. Los resultados son exhibidos en en el Capítulo 5 y su discusión se presenta en el Capítulo 6. Finalmente en el Capítulo 7 se desarrolla la conclusión y comentarios finales.

Capítulo 2

Mercados financieros

En este capítulo se hace una revisión acerca del concepto de mercado financiero y su clasificación según los diferentes instrumentos financieros. Se explica brevemente su funcionamiento, operación y regulación. La investigación propuesta en este trabajo se enfoca al mercado accionario o de valores, concretamente al sector financiero.

2.1. Definición y conceptos básicos

Un mercado financiero es en escencia, un espacio físico, virtual o ambos en donde se realizan intercambios de instrumentos financieros y se definen los volúmenes de operación y sus precios (Banco de México, 2020). Es decir, se trata de un foro en el que oferentes y demandantes intercambian instrumentos como acciones, derivados, tasas de interés, deudas, entre otros. Estos mercados surgieron en el siglo XVII ante la necesidad de actividades mercantiles (comercios y empresas).

En México, fue en el año de 1864 en el que surgió el primer banco. Se inaugura la primer sucursal del Banco de Londres, México y Sudamérica. Su finalidad era realizar operaciones mercantiles y bancarias como el descuento de letras de cambio, conceder préstamos a una tasa de interés y con la garantía de un bien; recibir depósitos de dinero, ahorros ofreciendo una tasa de interés atractiva para el público; apertura de cuentas corrientes, descuento de libranzas y negociación de letras de cambio sobre las principales ciudades de Europa, América y del país. Además, ofrecía la emisión y circulación de billetes, introduciendo así el uso del billete (Forbes, 2017).

Actualmente, el sistema financiero mexicano tiene un papel central en el desarrollo de la economía. Su función principal es intermediar la oferta y demanda de dinero. Al igual que en cualquier sistema financiero, participan diferentes intermediarios y organizaciones, no obstante, los bancos son los más importantes. Existen instituciones encargadas de regular el sistema financiero mexicano, entre ellas: la Secretaría de Hacienda y Crédito Público, Comisión Nacional Bancaria y de Valores, Comisión Nacional de Seguros y Finazas, Comisión Nacional para la Protección y Defensa de los usuarios de Servicios Financieros (CONDUSEF), cuya función principal es vigilar el buen funcionamiento del sistema financiero mexicano.

Dentro de las entidades operativas se encuentra la Bolsa Mexicana de Valores (BMV), Instituciones para el depósito de valores, Asociación Mexicana de Intermediarios Bursátiles (AMIB), entre otras, y son intermediarios, grupos financieros e inversionistas. Por ejemplo, la BMV opera por concesión de la Secretaría de Hacienda y Crédito Público (SHCP).

Por otro lado, ante la necesidad de aumentar la operación accionaria en México, surge en 2017 la Bolsa Institucional de Valores (BIVA). El 25 de julio de 2018, BIVA inicio formlamente su operacion como la

nueva Bolsa de Valores en Mexico, promoviendo la inclusion financiera y el fortalecimiento del mercado de valores mexicano (BIVA, 2020).

Las principales funciones de la BMV son:

i. Instalar los mecanismos de operación entre oferentes y demandantes.

ii. Vigilar el apego a las normas por parte de los involucrados.

iii. Certificar las cotizaciones en bolsa

iv. Garantizar el cumplimiento de acuerdos entre emisoras y accionistas.

Con relación a las normas aprobadas por la Bolsa Mexicana de Valores, que regulan la transacción de títulos se tiene la Ley del Mercado de valores, Ley del Banco de México, Ley de la Comisión Nacional Bancaria y de Valores, Reglamento de la Secretaría de Hacienda y Crédito Público, Ley de sociedades de inversión, entre otras.

Se denomina instrumento financiero a todo aquel contrato que deriva en un activo financiero de una empresa y, de forma coincidente en el tiempo, en un pasivo financiero o a un instrumento de patrimonio en otra empresa (BBVA, 2020). La existencia de los mercados financieros proporciona la posibilidad de elegir entre diferentes instrumentos financieros. Los mercados financieros se clasifican por los diferentes instrumentos financieros. Entre los más importantes están:

i. Mercado de deuda: Tanto empresas privadas y públicas, necesitan fuentes de financiamiento para realizar sus actividades, por lo que optan por conseguir recursos a través de préstamos bancarios o bien, emitiendo instrumentos de deuda. Naturalmente, el mercado de deuda es la infraestructura donde se emiten y negocian los instrumentos de deuda.

ii. Mercado cambiario o de divisas: En este mercado, se compran y venden las distintas monedas extranjeras. Gracias al mercado cambiario, tanto entidades financieras, como individuos son capaces de hacer dinero a través de la paridad de las distintas monedas. En otras palabras, se invierte en monedas extranjeras con motivos de cobertura o inclusive de apreciación del dinero.

iii. Mercado de derivados: Un derivado se define como un instrumento cuyo valor depende de un activo subyacente. Este activo puede ser una mercancia o instrumento, tales como metales, tasa de interés, divisas y acciones. La principal función de los derivados es servir de cobertura ante las fluctuaciones de precio de los subyacentes, por lo que se aplican preferentemente a portafolios accionarios, obligaciones contraídas a tasa variable, pagos o cobranzas en moneda extranjera a un determinado plazo, planeación de flujos de efectivo, entre otros (MexDer, 2017).

iv. Mercado accionario o de valores: Antes de definir este mercado y puesto que esta investigación estará enfocada en él, será vital definir el concepto de acción. Las acciones son títulos que garantizan que el tenedor posee una parte de una empresa emisora. Es decir, se trata de una parte o fracción del capital social de una empresa o sociedad constituida como tal (Santander México, 2020). Formalmente, el mercado de valores es el conjunto de mecanismos que permiten realizar la emisión, colocación y distribución de los valores inscritos en el Registro Nacional de Valores e Intermediarios y aprobados por la Bolsa Mexicana de Valores (BMV) (Dias, 2005). Volviendo a la definición de mercado, la oferta está representada por los títulos emitidos por la BMV, mientras que la demanda son aquellos fondos disponibles para inversión procedentes de personas físicas o morales.

Las empresas emiten acciones puesto que es un medio de financiación alternativo a un préstamo bancario o emisión de deuda. En un préstamo de cualquier índole, la organización se ve obligada a pagar intereses y está sujeta a los términos impuestos por el prestatario. De esta manera, una empresa emite acciones con el objetivo de obtener efectivo a través de los títulos que se compren en el mercado. Dependiendo de los intereses de los compradores existen diversos tipos de acciones, las cuales son:

i. Acciones clásicas: Derechos idénticos de voto y beneficios para los accionistas.

ii. Acción con derecho a voto privilegiado: Derecho de voto doble para accionistas que posean las acciones por varios años.

iii. Acción con bono de suscripción: El poseedor puede suscribir/comprar una acción posteriormente a un precio convenido previamente utilizando el bono.

iv. Acción con dividendo prioritario: Sin derecho de voto sino dividendos otorgados en prioridad a ciertos inversionistas importantes de la empresa.

Asimismo, un accionista puede generar un beneficio a través de acciones que ofrezcan dividendos o bien ocupar alguna estrategia de comprar una acción barata y esperar el momento en que la acción suba para obtener un rendimiento positivo.

Naturalmente, el mercado de valores representa un gran motor de la economía de un país puesto que promueve que empresas grandes sigan obteniendo fondos para operar y seguir ofreciendo empleos, generando demanda, etc. Por lo tanto, el comportamiento de la bolsa de valores es un parámetro a seguir para conocer el dinamismo de una economía. A pesar de que una economía sea sana, las acciones que se emiten en ella se están operando con una demanda y oferta considerable. Por esta razón, se define el concepto de índice bursátil como el promedio ponderado de los precios de una lista específica de acciones durante un tiempo determinado (Low, 2015). Es decir, los índices bursátiles son útiles para conocer el comportamiento general de un portafolio de acciones. En México, la BMV publica el Índice de Precios y Cotizaciones (IPyC), que es el principal indicador de rendimiento del mercado accionario mexicano con base en las variaciones de precios de la muestra representativa que lo compone y del grupo de acciones que cotizan en la bolsa (BMV, 2019).

Para 2019, 143 emisoras cotizaban en la BMV. De acuerdo con su capitalización bursátil entre las principales emisoras se encuentran (El Economista, 2019):

i. Walmart de México (WALMEX), 1 billón de pesos.

ii. América Móvil (AMXL), 968,937 millones de pesos.

iii. Fomento Económico Mexicano (FEMSA), 637,998 millones de pesos.

iv. Grupo México (GMEXICOB), 357,098 millones de pesos.

v. Grupo Financiero Banorte (GFNORTEO), 306,194 millones de pesos.

Dada la actividad y el giro de las diferentes emisoras, la Bolsa Mexicana de Valores clasifica a las empresas dentro de los siguientes grupos:

i. Energía

ii. Materiales

iii. Industrial

iv. Servicios y bienes de consumo no básico

v. Productos de consumo frecuente

vi. Salud

vii. Servicios Financieros

viii. Tecnología de la información

ix. Servicios de telecomunicaciones

x. Servicios públicos

Esta investigación se centrará en el sector financiero a fin de mostrar que una acción de este sector responde favorable o desfavorablemente a movimientos de precios de acciones del mismo sector (Figura 2.1). Principalmente porque en este sector cotizan las acciones de bancos, grupos financieros, aseguradoras, entre otras emisoras que mueven un gran volumen de capital dado que un gran porcentaje de mexicanos tienen participación en este sector ya que el 68 % de los mexicanos cuentan con al menos una cuenta bancaria (El Financiero, 2018).

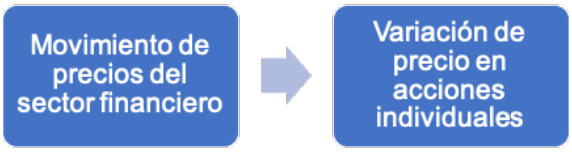

Figura 2.1: Reacción de una acción ante el mercado

Con esta propuesta se pretende predecir la variación futura de una acción dado el movimiento del mercado en el pasado. Para efectos ilustrativos, se eligió una entidad bancaria antigua en México para predecir la variación de precios de las acciones: Santander. Esta institución bancaria emite como Banco Santander (México), S.A., Institución de Banca Múltiple y Grupo Financiero Santander desde 2008 (Diario Oficial de la Federación, 2008).

2.2. Santander México

2.2.1. Historia y contexto en México

Santander México es uno de los bancos más antiguos en la historia del país. Surge en la creación del Banco de Londres, México y Sudamérica en el año 1864. El 22 de septiembre de 1932 nace el banco mexicano. Para 1958 se fusionan Banco Mexicano y el Banco Español, nacido para atender las necesidades de una amplia generación de empresarios españoles en México. Antes de ser Banco Español, tenía como nombre Banco Fiduciario.

Para 1990 se reformó la Constitución Política de los Estados Unidos Mexicanos para permitir la reprivatización total de los bancos comerciales mexicanos y el gobierno mexicano promulgó en 1991 la Ley de Instituciones de Crédito que llevó a la reprivatización de dichos bancos. Como parte de este proceso de privatización bancaria en 1992, Grupo InverMéxico adquirió Banco Mexicano Somex, que posteriormente adoptó el nombre de Banco Mexicano, S.A., Institución de Banca Múltiple y Grupo Financiero InverMéxico (Santander México, 2020).

El Banco Santander es una de las instituciones bancarias más importantes de México, alcanzando el tercer lugar del ranking de bancos con más activos del país (1,173,864 millones de pesos). Únicamente precedido por BBVA Banorte (1,249,817 millones de pesos) (Forbes, 2017).

Como se observa en la Figura 2.2, Banco Santander para 2017 se consolidó como uno de los bancos con mayores niveles de activos en el país. Por lo que se considera que es una de las instituciones bancarias más importantes del país cuya cotización de acciones genera un impacto significativo sobre el IPyC.

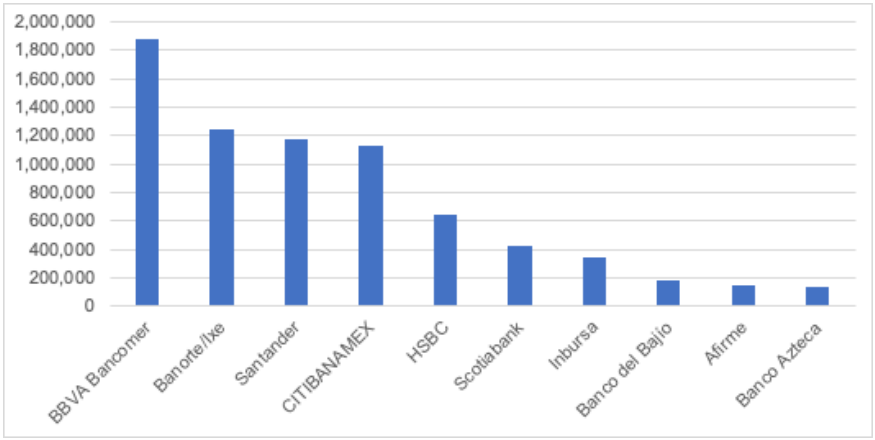

Figura 2.2: Top 10 bancos con mayores activos en México (en millones de pesos). Fuente: Elaboración propia con datos de CNBV.

2.2.2. Acciones de Santander México

Santander México comenzó a emitir acciones bajo la clave de pizarra BSMX el 15 de abril de 2008 en BMV, considerado un año complicado debido a la crisis financiera de 2008.

En la Figura 2.3 se muestra la cotización diaria de acciones de Santander México en el periodo del 4 de enero de 2016 al 4 de febrero de 2020 (1,037 registros). Es notable que la acción ha cotizado en un nivel no tan disperso, pues dentro del periodo ésta tuvo un nivel mínimo de $5.35 USD el 11 de febrero de 2016 y en un nivel máximo de $9.16 USD cuando la acción tuvo un valor medio de $6.93 USD.

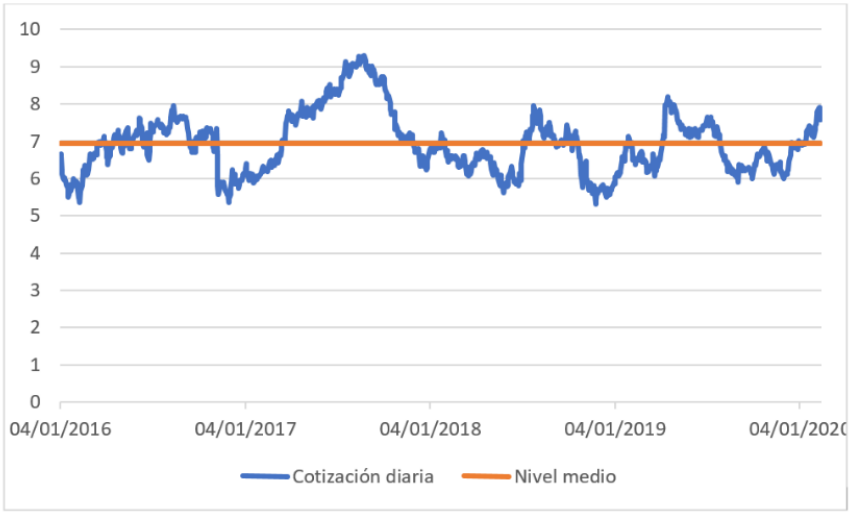

Figura 2.3: Serie de cotización diaria de Santander México y valor medio (Precios originalmente en dólares, estandarizados). Fuente: Elaboración propia con precios de Yahoo! Finance

El desarrollo metodológico de la investigación se centrará en tratar de predecir las variaciones de precio de la acción de Santander a través del rendimiento de acciones del mismo sector. Esto es, se pretende predecir si la acción de Santander subirá o bajará de precio dado el comportamiento del sector.

Capítulo 3

Modelos para el estudio de los mercados financieros

El interés general de las organizaciones o personas es predecir el comportamiento de ciertos instrumentos financieros a través del tiempo. Esto ha llevado a la búsqueda de diferentes métodos y herramientas matemáticas para conseguirlo. Por lo tanto, la necesidad de buscar patrones en la información histórica para poder predecir el futuro es un factor fundamental en este proceso. Considere el siguiente ejemplo, de manera empírica un inversionista puede percatarse que cierta acción del mercado tiende a subir cuando el tipo de cambio peso/dólar ha subido. Esto sugiere que, si en algún otro momento observa tal comportamiento en el mercado, el inversionista va a optar por vender todas sus acciones puesto que el mercado va al alza por el comportamiento del tipo de cambio.

Esta investigación pretende realizar un análisis similar al descrito en el párrafo anterior. Sin embargo, se busca basarse en información objetiva y datos históricos. Esto en beneficio de la acertada toma de decisiones basada en un criterio rígido y objetivo. Contrastando ambas partes, el inversionista del ejemplo basa sus decisiones en experiencias pasadas motivadas por su intuición. Mientras que la investigación se basa en comportamientos típicos en el pasado que estadísticamente pueden ser pronosticados.

3.1. La hipótesis de los mercados eficientes

La predicción de precios de acciones ha sido ampliamente estudiada desde diversos marcos metodológicos. En la década de 1950 se llevaron a cabo diversos esfuerzos de aplicación de cómputo para resolver problemas de economía. Varios de estos métodos se basan en series de tiempo, pues así como otros instrumentos financieros, la observación de los precios a través del tiempo debería mostrar tendencias o ciertas estacionalidades. Sin embargo, se llegó a la conclusión de que éstos eran igualmente probables de subir o bajar de precio. Este resultado sorprendió a los economistas puesto que parecía implicar que los mercados más allá de seguir ciertas tendencias, éstos reaccionan al pánico y no parecen ser controlados por cierta regla lógica. Más adelante, se vislumbró que los mercados no necesariamente han de ser racionales, pero sí son eficientes (Bodie, Kane y Marcus, 2011).

Un mercado eficiente es aquel en el que los precios reflejan toda la información disponible. En otras palabras, que el precio de mercado de un instrumento es un estimador insesgado de su verdadero valor. (Chen y Zheng, 2008). El concepto de eficiencia de los mercados es vital para modelar un problema de predicción, puesto que éste asegura que todo el público inversionista cuenta con la misma información, y el hecho de que la acción suba o baje depende de la existencia de nueva información acerca de la acción.

Un modelo de predicción utiliza información disponible como precios de cotización diario, noticias del mercado, etc., con el objetivo de hacer una estimación del precio futuro de una acción.

Suponga la siguiente situación: un investigador ha descubierto un modelo que predice el precio futuro de una acción. Si el investigador realiza cierta predicción y hace público que la acción subirá de $100 a $110, la agregación de esta información al mercado inevitablemente provocará que el precio de la acción suba y tome en algún momento dicho valor. Sin embargo, esto contradice la definición de un mercado eficiente puesto que el cambio de precio se vería influenciado por esta nueva información, por lo que el precio de la acción no estaba reflejando toda la información desde el principio.

De esta forma, el precio de una acción en el mercado es impredecible y completamente aleatorio dado que lo que hace que éste cambie es la aparición de nueva información que también es aleatoria. De cualquier modo, vale la pena realizar un esfuerzo para modelar al mercado a partir de una técnica que considere la mayor cantidad posible de información, tanto histórica como actual. La presente investigación integra información histórica. Sin embargo, existen grados de eficiencia del mercado en términos del tipo de información está disponible.

3.1.1. Forma débil de los mercados eficientes

Esta forma argumenta que los precios del mercado reflejan la información que puede ser obtenida a través de datos históricos. Bajo esta teoría analizar tendencias es inútil, pues si los inversionistas descubrieran ventajas a través del análisis de información histórica la utilizarían a su favor para obtener rendimientos al comprar/vender acciones (Bodie, Kane y Marcus, 2011).

3.1.2. Forma semi-fuerte de los mercados eficientes

Bajo esta forma los precios del mercado reflejan tanto información histórica como información fundamental sobre la emisora, tal como: estados financieros, hojas de balance, estados de flujo de efectivo, etc.

3.1.3. Forma fuerte de los mercados eficientes

Bajo esta forma, los precios del mercado reflejan tanto información histórica como información fundamental sobre la emisora, tal como: estados financieros, hojas de balance, estados de flujo de efectivo, etc.

En esta forma la información pública y la información clasificada (a la que sólo unos cuantos agentes dentro de las emisoras tienen acceso) debe reflejarse en el precio de la acción. De la hipótesis de los mercados eficientes se puede añadir que es difícil encontrar un mercado fuertemente eficiente pues la mayoría de los agentes del mercado sólo tiene acceso a información histórica pero ésta no es una forma adecuada de predecir precios.

Por último, esta forma de abordar el comportamiento del mercado pretende mostrar que, si bien los mercados no son racionales, éstos sí son eficientes pues solamente tres tipos de información pueden empujar el precio de una acción: información histórica, información financiera (fundamental) e información no pública de las emisoras.

3.2. Modelos clásicos para el mercado de valores

3.2.1. El análisis técnico

A pesar de que la hipótesis de los mercados eficientes no deja demasiada motivación para el análisis de información histórica, vale la pena abordar el mercado a través de tendencias y ciclos del mercado. El análisis técnico utiliza un conjunto de técnicas con el objetivo de encontrar patrones predecibles

en el precio de las acciones. También llamado en ocasiones análisis chartista, es utilizado en pro de realizar un análisis gráfico de la cotización histórica de una acción. De esta manera, el análisis técnico no considera que el precio justo de una acción sea su precio teórico pues el único precio justo para una acción está determinado por las leyes de oferta y demanda, es decir el único precio justo es aquel que el mercado esté dispuesto a pagar.

El análisis técnico considera básicamente tres premisas:

i. Todo lo que puede afectar el valor de un precio está descontado.

ii. Los precios del mercado siguen tendencias.

iii. El mercado tiene memoria. (Canelles, 2014)

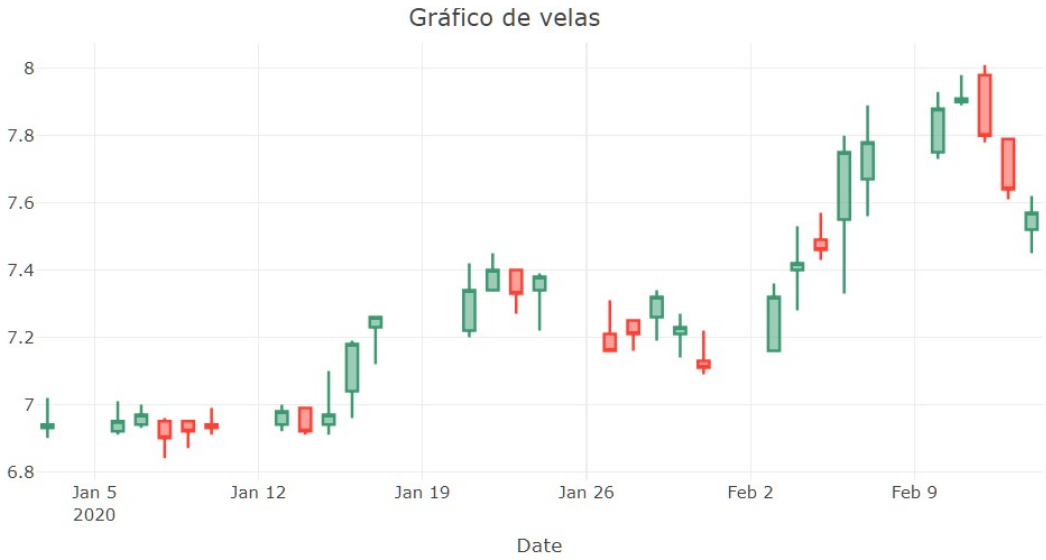

Figura 3.1: Gráfico de velas de Santander para Enero – Febrero 2019 (Precios en dólares). Elaboración propia con precios de Yahoo Finance.

Uno de los *charts* más característicos de el análisis técnico son los gráficos de velas (Padro et al, 2013). Estos proporcionanun resumen del comportamiento de una acción durante un día de cotización en específico. El color de la vela depende si el precio de cierre es mayor que el de apertura, en este caso la vela es de color verde o de color naranja en caso contrario. Además, las líneas verticales indican de abajo hacia arriba, el precio mínimo y máximo de la acción. La caja indica de abajo hacia arriba, el precio de apertura y el precio de cierre. De esta manera en la Figura 3.1, se puede inferir por ejemplo que cuando el precio de la acción comienza a subir ya ha seguido una trayectoria de velas verdes. Cuando el precio comienza a bajar, el color de las velas se torna naranja.

3.2.2. Análisis fundamental

A través del análisis fundamental se reúne la mayor cantidad de información posible con el objetivo de obtener un estimado del valor de sus acciones. De esta manera el análisis fundamental supone que los mercados no son eficientes pues el valor observado en el mercado no es representativo de su verdadero valor (Canelles, 2014).

Naturalmente, el análisis fundamental consiste en la conclusión que el accionista pueda sacar de estados financieros, balances, estados de resultados, valuaciones de activos, etc. Dado que no existe una forma

completamente objetiva de valorar una acción, el precio de éstas según el análisis técnico y fundamental, difieren en la mayoría de los casos. El análisis técnico basa sus conclusiones en análisis de ciclos y tendencias visibles en las gráficas de precios mientras que, el análisis fundamental asume que el precio de mercado observado de una acción no es un estimador insesgado y que la mejor estimación que se puede hacer de éste es a través del estudio y evaluación del desempeño financiero de la empresa emisora.

Desde el análisis fundamental existen dos perspectivas no excluyentes entre sí para la estimación del precio de una acción:

i. *Top down*

ii. *Bottom up*

El análisis *Top down* consiste en analizar el desempeño de una emisora de arriba hacia abajo, es decir observar factores externos al ejercicio operativo de la organización. Dentro de este rubro cabe el análisis de la economía global, partiendo desde cifras como producto interno bruto (P.I.B.), inflación, tasas de interés,etc. Tal entendimiento se finaliza con cifras globales del ejercicio financiero de la empresa emisora. Este análisis trata en general tres rubros: análisis de la economía, análisis sectorial y la compañía (Dias, 2005).

Por otro lado, el análisis *bottom up* parte primero de observar el desempeño de la emisora en términos financieros para luego involucrar agentes externos que influyen en el precio de sus acciones.

Volviendo a la teoría de los mercados eficientes, el análisis técnico y el análisis fundamental no logran realizar una predicción exacta del precio de una acción, puesto que ninguno de los dos es capaz de utilizar toda la información disponible. Sin embargo, se pueden realizar esfuerzos consolidando ambos análisis para obtener una estimación más precisa del verdadero valor de una acción.

Esta investigación se centra en analizar las variaciones de precio de una acción en particular en términos del comportamiento de acciones del mismo sector.

3.3. Modelos modernos para el estudio del mercado

Ante la necesidad de tratar de hacer predicciones sobre el precio de una acción surge la dificultad de analizar un gran volumen de información simultáneamente. Para tal propósito, se sugiere introducir *machine learning* (aprendizaje de máquina). El aprendizaje de máquina, es una rama de la computación que intenta hacer que la máquina aprenda.

Por muy obvio que esto parezca, esto tiene sentido puesto que el investigador contando con ciertos datos que pueden ser imágenes, texto, sonido, bases de datos, entre otras, diseña y aplica algoritmos de tal manera que, con base en lo aprendido, la máquina aprende a clasificar, predecir y proyectar.

Existen dos tipos de aprendizaje: supervisado y no supervisado. El aprendizaje supervisado parte de una serie de observaciones x_i y una variable objetivo t, escritos como un conjunto de datos x_i, t_i para $i = 1, \ldots, N$. Tal pareja de vectores funge como *dataset* de entramiento (conjunto de datos de entramiento).

En otras palabras, a través de las observaciones x_i, t_i, se aprende de la población y se ejecuta el aprendizaje en nuevos individuos para diferentes propósitos. El problema de la investigación recae sobre este tipo de aprendizaje, puesto que los precios de las acciones son conocidos, así como la variable objetivo, la cual es binaria (sube o baja).

Por otro lado, el aprendizaje no supervisado no cuenta con las clasificaciones de cada observación por lo que busca patrones en las variables de entrada, de tal manera que sea posible realizar una distinción clara entre categorías.

Para contrastar los conceptos anteriormente mencionados, considere el siguiente ejemplo. Un banco desea predecir si un cliente va a caer en incumplimiento de pago, tomando en cuenta la edad, sexo, antigüedad, monto promedio de pago (variables de entrada) y la clasificación como moroso o no moroso (variable objetivo). Eventualmente, la compañía aplicaría cierto algoritmo con la finalidad de aprender del conjunto de datos históricos y posteriormente clasificar a los individuos como morosos o no morosos, dadas sus características de entrada. Este tipo de aprendizaje es supervisado, puesto que el banco conoce de por medio cuáles son las características de un moroso y se aplica un modelo para clasificar a un moroso que no pertenece al conjunto de datos inicial.

Dentro del aprendizaje supervisado existen básicamente dos métodos fundamentales: regresión y clasificación. Tales métodos serán desarrollados a profundidad más adelante. En contraste, tomando al mismo banco como ejemplo, suponga que se pretende clasificar a los clientes como buenos, regulares o malos. Sin embargo, *a priori* el banco no tiene conocimiento de qué caracteriza a un buen cliente, a uno regular o a uno malo. Entonces parte de una base de datos de ejemplo para buscar patrones dentro de las variables de tal manera que se aprende de los datos y se llega a un modelo que clasifique a los clientes bajo esas categorías (Smola, 2008).

3.3.1. Introducción al Aprendizaje Supervisado

Como se ha comentado anteriormente, el aprendizaje de máquina puede ser dividido en dos paradigmas disitintos: supervisado y no supervisado. El aprendizaje supervisado se puede definir como el paradigma que busca una relación entre el *input* (también llamado entrada o aprendizaje) y el *output* (también llamado variable objetivo) (Liu y Wu, 2012). Es decir, a través de una estructura matemática pretende encontrar una función que reproduzca un resultado (predicción), dado un *input*.

También, dentro del aprendizaje supervisado se pueden encontrar dos tipos de modelos: regresión y clasificación.

Los modelos de regresión son un conjunto de técnicas estadísticas que tienen como objetivo obtener inferencias acerca de relaciones funcionales entre variables endógenas y variables exógenas (Goldberg y Cho, 2010).

Entre los modelos de regresión más utilizados, se pueden encontrar la regresión múltiple por mínimos cuadrados y regresión logística.

Por otro lado, los modelos de clasificación pretenden generalizar una estructura conocida con el fin de aplicarlo a nuevos datos. Es decir, una vez encontrada esta estructura se buscará aplicarla para nuevos conjuntos de datos donde no se conoce su clasificación.

Entre los algoritmos más conocidos para clasificación están: MVS (Máquinas de Soporte Vectorial), árboles de decisión, k vecinos más cercanos, el modelo ingenuo de Bayes, entre otros (Hiba Satia et al, 2019).

Un ejercicio interesante aplicando el modelo de MVS lo presentan Hiba Satia *et al.* (2019), para predecir la dirección (sube o baja) basado en información histórica de la acción. El presente trabajo como ya se ha abordado, utilizará la regresión logística para predecir la dirección de el precio de una acción, sujeta a los cambios o comportamiento del mercado financiero de valores.

3.3.2. Introducción al Aprendizaje No Supervisado

A diferencia de los métodos supervisados, los métodos no supervisados no cuentan con una clasificación *a priori* de los datos. Por lo tanto, únicamente buscan encontrar patrones subyacentes en la base de datos de ejemplo. Entre los métodos más conocidos de modelos no supervisados está la teoría de *clusters*, componentes principales y análisis factorial (Nasteski, 2017).

(Nanda, Mahanti, Tiwari, 2010), hacen un interesante análisis de *clusters* en el mercado de valores para encontrar portafolios de acciones con características en común. Lo cual es un modelo apropiado para tal propósito, pues antes de construir un portafolio de acciones, no se cuenta con etiquetas o clasificaciones para su uso, tales como bueno, malo, eficiente, no eficiente, etc. Por lo tanto, se contruye un modelo de *clusters* para encontrar aglomeraciones (o portafolios) que tienen características heterogéneas entre sí.

Por otro lado, el análisis de componentes principales ha sido ampliamente utilizado en análisis exploratorios, puesto que la reducción de dimensión otorga información valiosa acerca de la base de ejemplos, sin que esta información sea visible a través de métricas tradicionales.

Capítulo 4

Desarrollo metodológico

En el Capítulo 3 se abordaron dos modelos de aprendizaje de máquina muy importantes para el desarrollo de la investigación. Por un lado, por su definición la regresión lineal servirá como apoyo para encontrar a través de un análisis exploratorio patrones latentes dentro del portafolio conformado por los precios de las acciones. Posteriormente, se definirán los parámetros con los que se aplicará el modelo de regresión logística con motivo de predecir la variación del precio de una acción sujeto a las variaciones del mercado.

4.1. El sector financiero en la Bolsa Mexicana de Valores

Dentro del sector financiero en la bolsa se enlistan 27 empresas. Sin embargo, esta clasificación incluye bancos, bienes inmobiliarios, grupos financieros, intermediarios financieros no bancarios, mercados financieros, organizaciones auxiliares de crédito, seguros, servicios financieros diversificados, siefores y sofoles (BMV, 2020). Sin embargo, esta clasificación es bastante extensa y general, por lo que se opta por únicamente analizar emisoras del subsector de bancos y grupos financieros.

En el Cuadro 4.1 se listan las emisoras cuyos precios de acciones se analizarán a lo largo de la investigación. Es decir, de un total de 27 emisoras en total en el sector financiero queda un remanente de 8 emisoras únicamente de los subsectores bancos y grupos financieros.

Clave	Razón Social	Subsector
BBAJIO	BANCO DEL BAJÍO, S.A., INSTITUCIÓN DE BANCA MÚLTIPLE	Banco
BBVA	BANCO BILBAO VIZCAYA ARGENTARIA, S.A.	Banco
BSMX	BANCO SANTANDER MEXICO, S.A., INSTITUCIÓN DE BANCA MULTIPLE, GRUPO FINANCIERO SANTANDER	Banco
GFINBUR	GRUPO FINANCIERO INBURSA, S.A.B. DE C.V.	Grupo financiero
GFMULTI	GRUPO FINANCIERO MULTIVA S.A.B. DE C.V.	Grupo financiero
GFNORTE	GRUPO FINANCIERO BANORTE, S.A.B DE C.V.	Grupo financiero
GPROFUT	GRUPO PROFUTURO, S.A.B. DE C.V.	Grupo financiero
VALUEGF	VALUE GRUPO FINANCIERO, S.A.B. DE C.V.	Grupo financiero

Cuadro 4.1: Emisoras en el Sector Financiero. Fuente: BMV

4.1.1. Análisis exploratorio sobre el sector financiero

Como análisis exploratorio, se toma en cuenta la cotización diaria de las acciones del sector financiero y subsector bancario (con el precio de cierre). El precio de cierre de una acción es el precio con el que una emisora culminó el día de cotización.

En la Figura 4.1 , se muestra la evolución de la cotización de las acciones de Santander, BBVA y Banbajío. Es notorio que las acciones de BBVA han sostenido una tendencia a la baja desde el 8 de junio de 2017. Por otro lado, las acciones de banbajio durante el periodo del 08/06/2017 al 25/10/2018 mantuvo una tendencia al alza, y después de este periodo comenzó una etapa de decrecimiento. Finalmente, las acciones de Santander comenzaron con una tendencia a la baja en el periodo 29/08/2017 hasta el 27/11/2018. Es importante mencionar que de este grupo de acciones, Santander mantiene una tendencia al alza significativa desde el 19/12/2019 hasta el 14/02/2020.

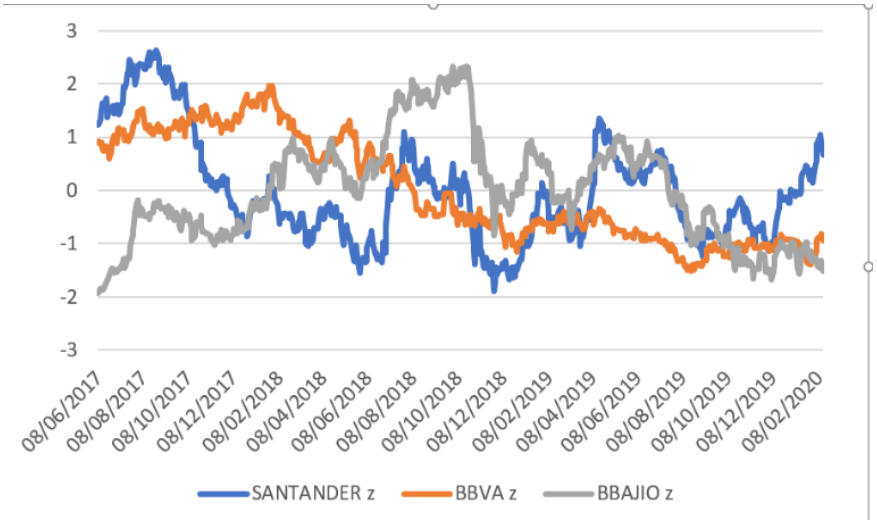

Figura 4.1: Serie de cotización diaria del subsector bancario (precios en pesos estandarizados). Fuente: Elaboración propia con precios de Yahoo! Finance.

Un parámetro para tomar en cuenta en el análisis del mercado es el rendimiento que genera un activo financiero. En este caso, se analizan los rendimientos diarios que generan las acciones del subsector bancario. Este rendimiento se calcula como:

$$Rendimiento_t = \ln\left(\frac{\text{Precio de cotización}_{t+1}}{\text{Precio de cotización}_t}\right).$$

En el Cuadro 4.2 se muestran las estadísticas descriptivas de los rendimientos de las acciones del subsector bancario. De esta tabla, se puede inferir que Santander es la acción que ha mostrado un mejor rendimiento a lo largo del periodo (11.69 %). Mientras que la acción de banbajío mostró el rendimiento más bajo respecto a las demás (-9.13 %). Es de suma importancia tener en cuenta que estas métricas son útiles para contrastar el comportamiento del portafolio, pero de ninguna manera se deben tomar como conclusiones, pues por ejemplo, los valores mínimo y máximo fueron captados únicamente en una observación a lo largo del tiempo.

En la Figura 4.2 se puede inferir que los rendimientos de los precios siguen una distribución acampanada centrada en la media igual a cero. Esto se interpreta como que a pesar de que la acción de Santander mostró un comportamiento atípico positivo (rendimiento del 11.69 %) y que la acción de

Métrica	R SANTANDER	R BBVA	R BBAJIO
Mínimo	-0.0875	-0.0690	-0.0913
Máximo	0.1169	0.0534	0.0582
Desviación estándar	0.0190	0.0164	0.0150
Media	-0.0001	-0.0005	0.0001

Cuadro 4.2: Estadísticas descriptivas de los rendimientos de las acciones. Fuente: Elaboración propia

banbajío mostró un rendimiento atípico negativo (-9.13 %), el rendimiento de las acciones sigue una distribución aproximadamente normal centrada en cero. Lo cual implica que los rendimiento siempre siguen comportamiento típico, a pesar de observaciones extremas.

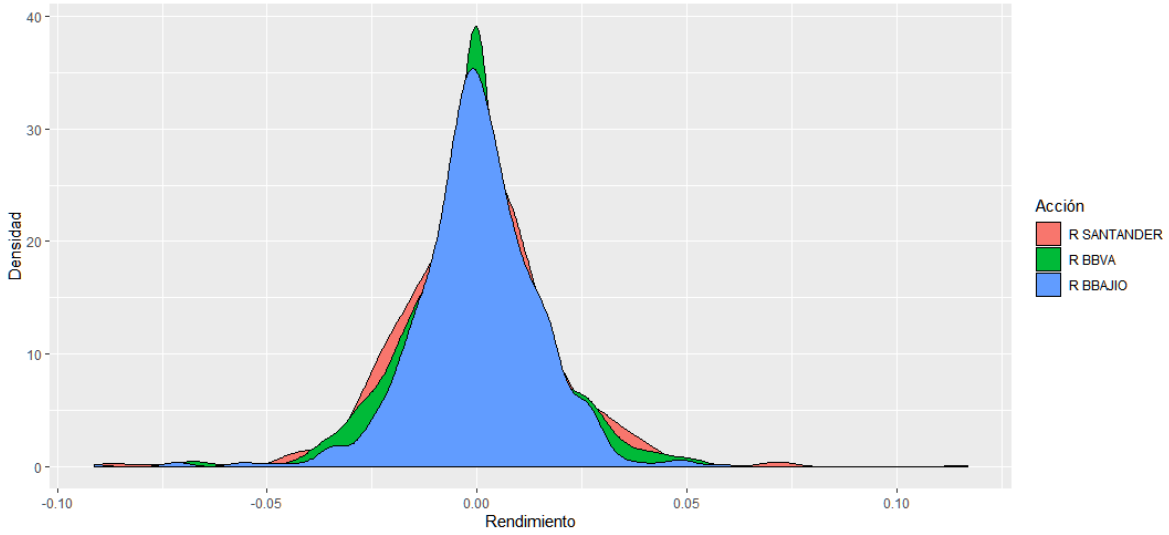

Figura 4.2: Densidad de los rendimientos. Fuente: Elaboración propia con precios de Yahoo! Finance

A lo largo de la sección, también se muestran visualizaciones (gráficas) con el objetivo de obtener un mejor entendimiento del problema que complemente la aplicación de los modelos.

4.2. Datos y análisis exploratorio

Utilizando los precios de cierre de cotización de las emisoras mencionadas en un periodo del 24 de mayo de 2016 al 1 de junio de 2020, deja un total de 660 registros para la aplicación de los modelos ($n = 660$). Con este tamaño de muestra se analiza a través de visualizaciones de patrones que puedan ser útiles para el entendimiento del problema (predecir la variación de futura de una acción en particular).

4.2.1. Visualizaciones de la muestra

Partiendo con el tamaño de muestra n se mantienen únicamente registros en los cuales, las 8 emisoras del sector financiero hayan emitido precios de cierre de manera unánime, puesto que en ocasiones no hay precios emitidos por motivo de días feriados, contingencias, etc., dejando un total de 692 registros.

Primeramente, es necesario visualizar las distribuciones de los precios de cotización de cada uno de las emisoras con el objetivo de comparar el comportamiento de sus precios.

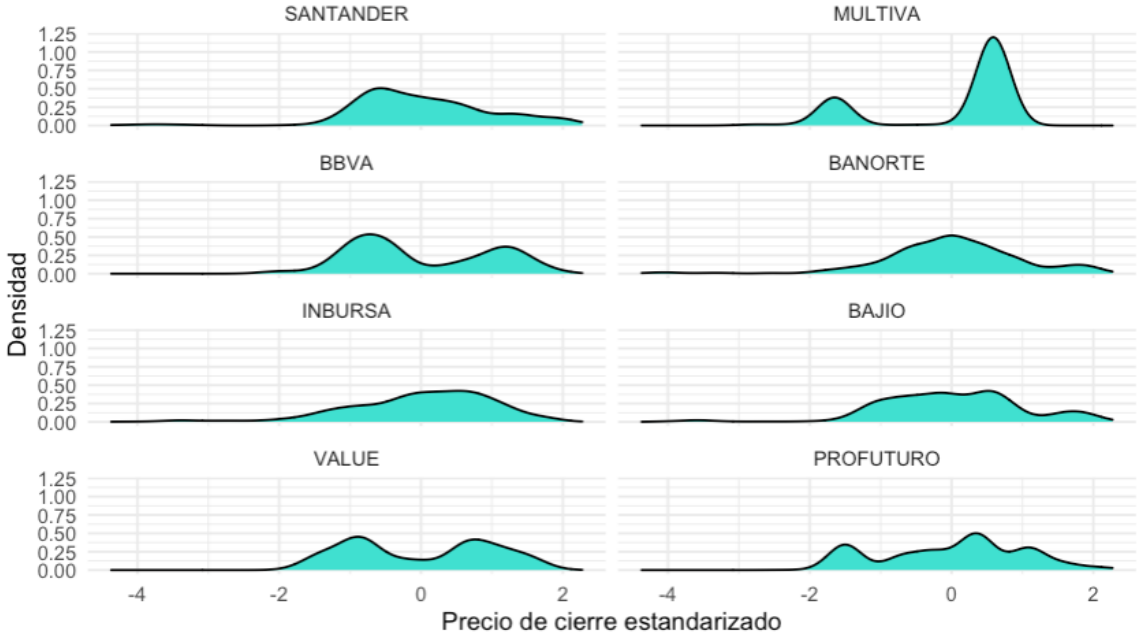

Figura 4.3: Densidad de los precios de cierre estandarizados. Fuente: Elaboración propia con precios de Yahoo! Finance.

En la Figura 4.3, se muestra la función de densidad empírica de cada uno de los precios de cotización de las 8 emisoras del sector financiero (estandarizados). Es importante destacar que de esta gráfica se puede inferir que todas las emisoras tienen distribuciones de probabilidad muy diferentes en términos de sus precios de cierre. Esto implica que en la práctica es atípico observar precios de cotización similares o que estén centrados en algún punto en particular.

Como se comentó en el Capítulo 1 y 2, el único precio justo en para una acción determinada es aquel que el mercado esté dispuesto a pagar. De aquí se puede ver que claramente hay acciones que están mejor cotizadas en el mercado (mantienen un precio de cotización mayor a las demás)

Complementando el argumento anterior, en la Figura 4.4, el precio mediano de cotización (punto rojo) de las emisoras Multiva y Profuturo es visiblemente mayor que los del resto. Lo que podría sugerir que estas últimas acciones tendrán poco efecto en el comportamiento del sector financiero (sus precios de cotización no se mueven a la misma velocidad que los del resto).

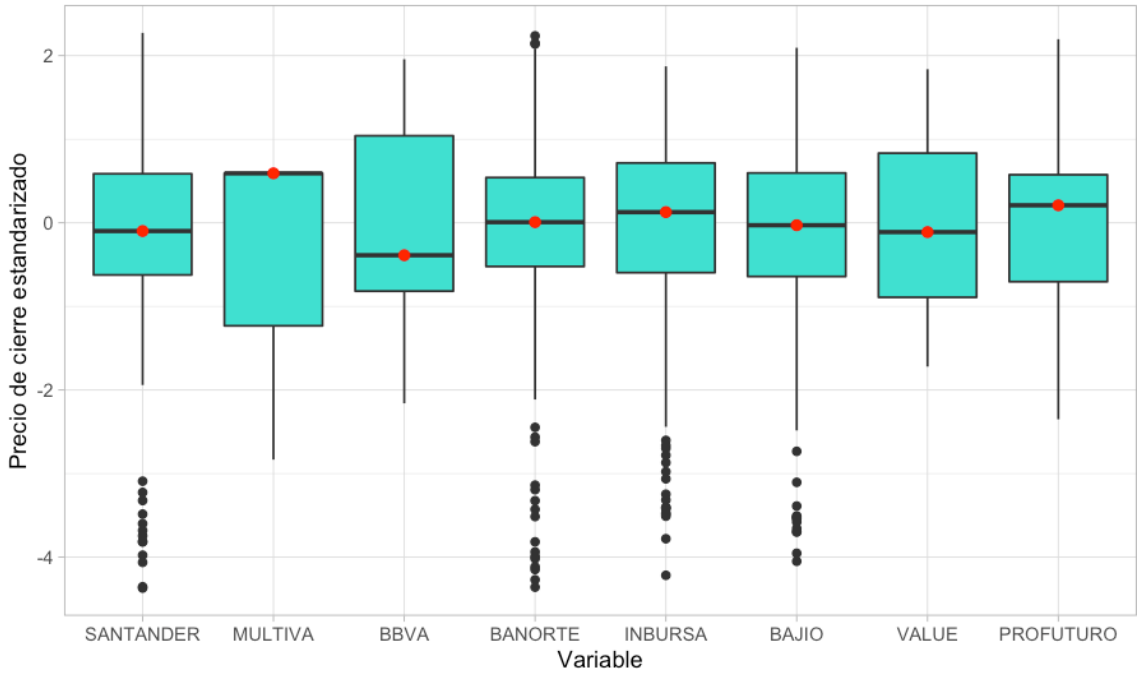

Figura 4.4: Diagrama de violín de los precios de cierre por emisora. Fuente: Elaboración propia con precios de Yahoo! Finance

En la Figura 4.5 se observa la relación lineal entre cada par de emisoras. Puesto que la investigación está centrada en predecir la variación futura de las acciones de Santander, es importante resaltar que esta acción guarda una correlación positiva fuerte con las acciones de Banorte. Mientras que sigue una correlación leve con acciones de BBVA, Inbursa y Bajío. Por otro lado, ésta guarda una correlación negativa leve con acciones de Multiva.

De igual manera, vale la pena destacar que de todas las combinaciones, las que más llaman la atención, son las correlaciones entre Inbursa y BBVA (0.85), Profuturo y Multiva (0.77) así como BBVA y Multiva (-0.76).

De las Figuras 4.3, 4.4 y 4.5, se puede inferir que todas los precios de las acciones del portafolio siguen distribuciones de probabilidad diferentes (por lo tanto, el mercado no las demanda con la misma intensidad). También, que existen acciones que reciben mejor atención por parte del mercado (de ahí su precio relativamente elevado) y por último, que es posible asociar el incremento o decremento de acciones de Santander con el comportamiento de otras acciones del sector.

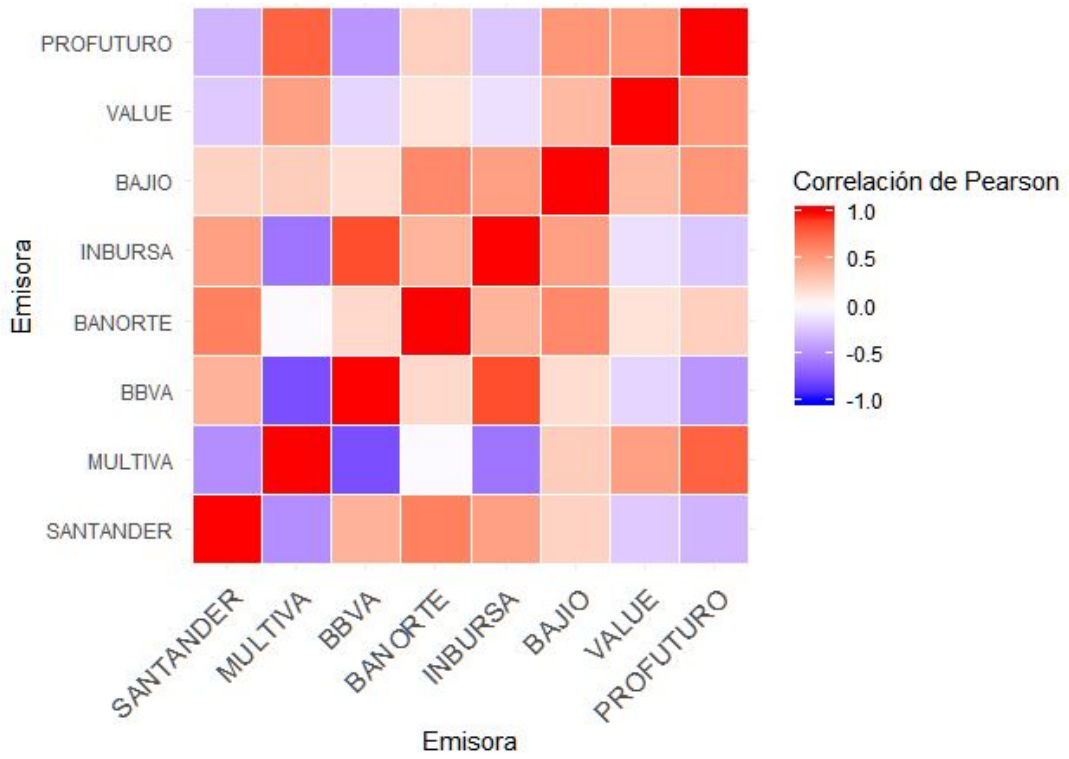

Figura 4.5: Mapa de calor de matriz de correlación de los precios. Fuente: Elaboración propia con precios de Yahoo! Finance

4.3. Métodos

En esta sección de abordarán dos métodos: regresión lineal y logística. Los métodos de regresión se usan ampliamente para analizar la relación entre una variable dependiente y una o más variables independientes. El método de regresión más popular es la regresión lineal que utiliza el método de mínimos cuadrados. Sin embargo, es aplicable si la variable dependiente es continua, independiente e idénticamente distribuida solamente. En los casos en que la variable dependiente es categórica, el análisis de regresión lineal no es apropiado.

4.3.1. Coeficiente de correlación muestral

Para efectos de inferencia estadística, se ha apelado a utilizar métricas que describan la relación o correlación latente entre dos variables numéricas. Para tal propósito se utilizó el coeficiente de correlación de Pearson, que se utiliza frecuentemente para medir la intensidad de la relación lineal que existe entre dos variables.

Sea $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ una muestra aleatoria bivariada de tamaño n. El coeficiente de correlación muestral se define como: (Wackerly, Mendenhall y Scheaffer, 2008):

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

El coeficiente de Pearson puede escrbir en términos de la desviación estándar, esto es:

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

Por la desigualdad de Cauchy-Schwartz, se puede mostrar que:

$$|r| \leq 1.$$

Un aspecto importante acerca de esta métrica es que a partir de la desigualdad mencionada, se pueden tomar ciertas inferencias, las cuales son:

Si $r \approx -1$, se argumenta que las variables x_i e y_i tienen correlación negativa, lo cual significa que cuando una aumenta su valor, la otra tiende a disminuir.

Si $r \approx 1$, se argumenta que las variables x_i e y_i tienen correlación positiva, lo cual significa que cuando una aumenta su valor, la otra tiende a aumentar también.

Si $r \approx 0$, se argumenta que las variables x_i e y_i tienen correlación nula, o que son no correlacionadas, lo cual implica que no existe evidencia estadística para argumentar que las variables tengan un efecto entre ellas.

En general se dice que el coeficiente de correlación de Pearson refleja la relación lineal entre dos variables sin supuestos *a priori*. (Restrepo y González, 2007)

4.3.2. Regresión lineal

La regresión es una método estadístico cuyo fin es analizar la relación entre una variable dependiente y una o varias variables independientes. Por ejemplo, es posible utilizar este método para estimar el peso de una persona (variable dependiente) utilizando como variables regresoras (variables dependientes) la presión arterial, pulso cardíaco, estatura, etc. Volviendo al tema de investigación, se busca poner al precio de una acción en función de otras variables endógenas, tales como tipo de cambio, movimiento del sector, comportamiento del mercado, tasa de fondeo gubernamental, etc.

Cabe resaltar que esta técnica es usada para calcular el precio de una acción, dados los parámetros mencionados anteriormente, aunque para fines de esta investigación, el análisis de regresión fungirá como discriminante ante variables poco significativas al precio del título.

En efecto, se define el modelo de regresión (Rawlings, 2005) como:

$$\boldsymbol{Y} = \beta_1 \boldsymbol{X}_1 + \beta_2 \boldsymbol{X}_2 + \cdots + \beta_m \boldsymbol{X}_m,$$

donde $\boldsymbol{Y}$ representa el precio de la acción, β_k representa al coeficiente del vector regresor $\mathbf{X_k}$ para toda $k = 1, 2, \ldots, m$ y m representa el número de variables independientes escogidas para el modelo. Esto es, se está colocando al precio de la acción en función de m parámetros a escoger. Posteriormente se deben de estimar los parámetros β_k como sigue:

Sean $x_1, x_2, \ldots, x_m$ las m variables independientes del modelo de regresión, suponiendo una muestra de n observaciones por cada atributo x_i.

Se forma el siguiente vector de observaciones de la variable dependiente $\boldsymbol{Y}$. Así como la matriz $\boldsymbol{X}$ de observaciones de las variables independientes x_k y el vector de coeficientes β_k para toda $k = 1, 2, \ldots, m$.

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} x_{1,1} & \cdots & x_{1,m} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,m} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_n \end{pmatrix}.$$

De tal forma que el modelo de regresión es de la forma

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \beta_1 x_{1,1} + \beta_2 x_{1,2} + \cdots + \beta_m x_{1,m} \\ \vdots \\ \beta_1 x_{n,1} + \beta_2 x_{n,2} + \cdots + \beta_m x_{n,m} \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}, \epsilon_i \sim N(0, \sigma^2).$$

En forma matricial:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

Se define $e_i = y_i - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_i$, es decir, la diferencia entre el valor de la observación i y la estimación i del modelo. Al minimizar la función

$$\begin{aligned} LS(\widehat{\boldsymbol{\beta}}) =& \sum_{i=1}^{n} e_i^2 \\ =& (\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^t \left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right) \end{aligned}$$

se obtiene que la función LS se minimiza con el estimador $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X^t X})^{-1}\boldsymbol{X^t Y}$ (Montgomery et al, 2002)

Volviendo al contexto de la investigación, la matriz $\boldsymbol{X}$ está compuesta por los vectores de observaciones de las variables independientes. Mientras que los precios de la acción de la emisora componen al vector $\boldsymbol{Y}$. Como se ha mencionado con anterioridad, el análisis de regresión se empleará únicamente como discriminante ante variables poco significativas respecto del precio de la acción.

Utilizando el hecho que

$$Var\left(\widehat{\boldsymbol{\beta}}\right) = \sigma^2(\boldsymbol{X}^t\boldsymbol{X})^{-1},$$

siendo $\widehat{\sigma^2} = \frac{\boldsymbol{e}^t\boldsymbol{e}}{n-k}$ un estimador insesgado de σ^2.

Sean $\boldsymbol{V} = \widehat{\sigma^2}(\boldsymbol{X}^t\boldsymbol{X})^{-1}$, $Se\left(\hat{\boldsymbol{\beta}}_i\right) = V_{i,i}$ para toda $i = 1, \ldots, m$. Es decir, $Se\left(\hat{\boldsymbol{\beta}}_i\right)$ representa cada elemento de la diagonal de la matriz $\boldsymbol{V}$ que es el error estándar para el estimador $\widehat{\boldsymbol{\beta}}_i$.

Para medir la significancia de las variables regresoras, se utilizan la prueba de hipótesis:

$$\begin{aligned} H_{0,k} :& \beta_k = 0, \\ H_{a,k} :& \beta_k \neq 0 \end{aligned}$$

Se define el estadístico de prueba bajo H_0:

$$\frac{\widehat{\boldsymbol{\beta}}_i}{Se\left(\hat{\boldsymbol{\beta}}_i\right)}.$$

Donde este cociente sigue una distribución T de Student con parámetros $n - m$ grados de libertad (t_{n-m}). Se rechaza H_0 si $\frac{\widehat{\boldsymbol{\beta}}_i}{Se\left(\hat{\boldsymbol{\beta}}_i\right)} > t_{\alpha,n-m}$, a un nivel de significancia α (McCullagh, 2000).

En otras palabras para la discriminación de variables, una variable independiente va a ser poco significativa si no se rechaza H_0.

Una vez obtenido un modelo de regresión lineal, como ya se ha discutido, es importante evaluar la significancia de sus parámetros a través de contrastes de hipótesis de nulidad. Sin embargo, además de tal análisis, se tiene que exhibir alguna métrica que refleje el nivel de ajuste de un modelo dado un conjunto de datos. Una métrica para tal propósito es el coeficiente R^2 (r cuadrada).

Antes de mostrar la definición R^2, se deben definir dos conceptos fundamentales, el de la suma cuadrada de los residuos y los residuos totales:

Sea $\hat{y_i}$ el i-ésimo valor ajustado para $i \leq n$, entonces se define la suma cuadrada de los residuos como:

$$SSR = \sum_{i=1}^{n} (y_i - \hat{y_i})^2$$

Por otro lado, la suma total de los residuos:

$$SST = \sum_{i=1}^{n} (\hat{y_i} - \bar{y})^2$$

Por la desigualdad de Cauchy-Schwartz, se puede mostrar que:

$$SSR \leq SST$$

Por lo que:

$$R^2 = \frac{SSR}{SST} \leq 1$$

(Bartels, 2005)

Dado que la suma cuadrada de los residuos está acotada por la suma total de los residuos (o varianza de la muestra), se puede decir que la R^2 se interpreta como el porcentaje de varianza explicada por el modelo.

Sin embargo, a medida que se aumenta el número de predictores (m), la R^2 tiende a incrementar su valor, por lo que se introduce otra métrica que pondera a la R^2 con el número de estimadores que utiliza el modelo. Se define la R^2 ajustada como:

$$R^2_{adj} = 1 - (\frac{n-1}{n-m-1})(1 - R^2)$$

(Karch, 2019)

Entre más cercano se encuentre este coeficiente a 1 (100 %), esto indica que el modelo escogido explica adecuadamente la varianza generada por la muestra.

4.3.3. Regresión logística

La regresión logística, como la regresión de mínimos cuadrados, es una técnica estadística que se utiliza para explorar la relación entre una variable dependiente y al menos una variable independiente. La diferencia es que, la regresión lineal se usa cuando la variable dependiente es continua, mientras que las técnicas de regresión logística se usan con variables dependientes categóricas. Aunado al hecho de que el análisis de regresión logística no requiera muchos supuestos lo hace preferible para su aplicación.

Dado que se desea conocer la probabilidad de alza o baja de Y, condicionada a los rendimientos de empresas del mismo sector, es natural hablar de probabilidad condicional, puesto que, para efectos de la investigación, al momento de predecir si una acción subirá o bajará de precio, tal probabilidad fungirá como umbral para decidir una respuesta positiva, por ejemplo si la probabilidad excede el 50 % se acepta que la acción subirá. En términos de probabilidad, lo anterior es:

$$P(Y|X_1, X_2, \ldots, X_k).$$

En otras palabras, se apela a modelar la variación de precio futura de la acción, condicionada a los rendimientos de empresas del mismo sector.

Un modelo de regresión logística se define como sigue (Dobson, 2008). Sea p la distribución condicionada $P(Y|X_1, X_2, \ldots, X_k)$ con valor esperado $E(Y|X_1, X_2, \ldots, X_k)$. Se tiene que:

$$\begin{aligned} \ln\left(\frac{p}{1-p}\right) &= g\left(E(Y|X_1, X_2, \ldots, X_k)\right) \\ &= \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k. \end{aligned}$$

En forma de vectores se puede considerar $\boldsymbol{X} = [1, X_1, X_2, ..., X_k]$ y $\boldsymbol{\beta}^t = [\beta_0, \beta_1, \beta_2, ..., \beta_k]$. Ahora, despejando a p de la ecuación se obtiene:

$$p(\boldsymbol{X}; \boldsymbol{\beta}) = \frac{1}{1 + e^{-\boldsymbol{X}\boldsymbol{\beta}}}$$

La función $\ln\left(\frac{p}{1-p}\right)$ es conocida también como *log-odds* y naturalmente $\frac{p}{1-p}$ como *odds* (en ocasiones también llamada momios). Lo cual es un modelo *ad hoc* para modelar la probabilidad condicional mencionada anteriormente. La estimación de parámetros se realiza a través de maximización de la función de verosimilitud utilizando métodos numéricos (McCullagh, 2000).

Se hará uso del lenguaje de programación *R* (específicamente el paquete glm) para calcular los parámetros, así como para definir el umbral y clasificar a una nueva base de datos. Posteriormente, se calculará la precisión de manera simple (casos correctamente clasificados/casos totales), así como la métrica ROC para medir la calidad de las predicciones del modelo.

En palabras llanas, tomando en cuenta el comportamiento del mercado en general en cierto periodo $t - h$ y la cotización del día siguiente de Santander a un tiempo t, se calculará la verosimilitud o probabilidad de que, dado cierto movimiento del mercado, la acción suba o baje de precio en el tiempo t. Estas probabilidades provendrán de un modelo de regresión logística.

4.4. Medidas de precisión y validación

La presente investigación pretende predecir la variación futura de una acción de Santander sujeta al comportamiento del mercado. Desde luego, hechas estas predicciones es de suma importancia validar su verosimilitud.

En la Sección 3.3 se definió el concepto de aprendizaje de máquina y sus dos categorías: aprendizaje supervisado y no supervisado. Dado que se aplicará un modelo de regresión logística que precisamente pertenece al tipo de métodos empleados en el aprendizaje supervisado, una vez entrenado el modelo se realizará un proceso de validación a través de la curva ROC (Receiver Operating Characteristic). Una vez obtenida esta curva, se calculará su área bajo la curva (AUC por sus siglas en inglés), teniendo una métrica que califique en gran medida la calidad de las predicciones. Por otro lado, en el análisis de regresión lineal, se presentan resultados de poner a las acciones de Santander en función de la cotización de otras acciones en el mercado, por lo que el concepto de R^2 también será definido en esta sección.

4.4.1. Matriz de confusión

Una matriz de confusión en un problema de clasificación binaria es una matriz cuyas entradas representan las cantidades de Verdaderos Positivos (VP), Falsos Positivos (FP), Falsos Negativos (FN) y Veraderos Negativos (VN)(Visa *et al.*, 2011). Es decir, en esta matriz se colocan el número de individuos correctamente clasificados por el modelo (VP y VN) y los no clasificados correctamente (FP,FN). Ver Matriz de confusión para clasificación binaria. (Matriz 4.4.1)

		Predicho	
		P	N
Observado	P	VP	FN
	N	F P	VN

(4.4.1)

4.4.2. Curva ROC

La curva ROC (por sus siglas en inglés) y su respectiva área bajo la curva AUC (por sus siglas en inglés) son métricas que permiten evaluar el rendimiento de un modelo de aprendizaje (Marsland, 2015).

La curva ROC es utilizada en el aprendizaje de máquina supervisado, particularmente cuando se tiene que clasificar individuos de una muestra en dos categorías. Esta curva describe el nivel de clasificación, dado un nivel de corte. En esta investigación, este corte está dado por la probabilidad de que una acción suba de precio.

Formalmente, se define a la curva ROC como:

$$ROC(c) : \begin{cases} y = S(c) \\ x = 1 - E(c) \end{cases}$$

donde la **sensibilidad** $S(c)$ es la tasa de positivos clasificados por el algoritmo correctamente (VP) como positivos, respecto al total de positivos observados (VP+FN). En términos de la matriz de confusión esto es:

$$S = \frac{VP}{VP + FN}.$$

Ahora, la **especificidad** $E(c)$ es la tasa de negativos clasificados correctamente por el algoritmo como negativos (VN), respecto al total de negativos observados (VN+FP). En términos de la matriz de

confusión esto es (Del Valle, 2015):

$$E = \frac{VN}{VN + FP}.$$

4.4.3. Área bajo la curva (AUC)

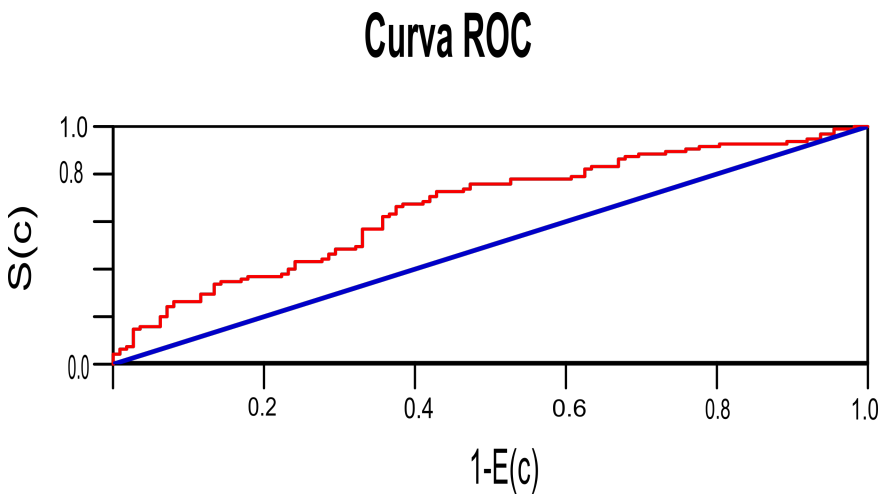

Figura 4.6: Curva ROC y función identidad. Fuente: Elaboración propia con datos simulados

Una vez obtenida la curva ROC, una forma ampliamente utilizada para validar la precisión y calidad de las predicciones, es el área bajo dicha curva (AUC). Esto pues esta área representa la probabilidad de que un individuo sea correctamente clasificado (como positivo o como negativo). Formalmente se define a la AUC como (Martínez-Camblor, 2007):

$$AUC = \int_0^1 ROC(t)\ dt.$$

En la Figura 4.6, se muestra la gráfica de la curva ROC en la cual se puede visualizar la comparación entre el complemento de la especificidad $(1 - E(c))$ y la sensibilidad. Por otro lado, la línea azul es la recta $y = x$. Una AUC cercana 1 naturalmente significa que el modelo está correctamente clasificando a los individuos como positivos o negativos.

Se puede probar que el área bajo la curva de la recta $y = x$ (línea azul) es exactamente 0.5, por lo que gráficamente una curva ROC cercana a la gráfica de la función identidad significa que la probabilidad de que clasifique adecuadamente a los individuos de una muestra es cercana a la de un volado. Por lo que se dice que la clasificación del modelo no es mejor que el azar.

De este modo, será un resultado deseable que las predicciones efectuadas por el modelo de regresión logística produzcan una curva ROC considerablemente por encima de la función identidad y que el área bajo la curva se lo más cercano posible a 1, pues se estaría garantizando que las acciones del mismo sector dictan en gran medida la variación de las acciones de Santander.

Capítulo 5

Resultados

En este capítulo se presentan los resultados del análisis de los resultados después de aplicar los métodos propuestos anteriormente. Esto es, se presentan los resultados del modelo de regresión lineal y del modelo de regresión logística a fin de discriminar variables poco significativas para determinar el precio de las acciones de Santander. Posteriormente, con las variables seleccionadas se aplicará el modelo de regresión logística con la premisa de entrenar un modelo de clasificación binaria y realizar una predicción futura.

Posteriormente se presentan las métricas de precisión y validación introducidas en la Sección 4.4.

5.1. Análisis de regresión lineal

Ahora interesa conocer la respuesta de las acciones de Santander ante los movimientos de todo el mercado. Es decir, encontrar un modelo que sea capaz de explicar algo como la siguiente relación:

$$Santander = f(sector).$$

Para lo cual, se apela a regresión lineal múltiple, poniendo a Santander como variable dependiente de las variables BBVA, Multiva, Banorte, Inbursa,Banbajío, Value, Profuturo. Sean los siguientes parámetros:

$$\begin{aligned} Y &: \text{Precio de cierre diario de Santander,} \\ X_1 &: \text{Precio de cierre diario de BBVA,} \\ X_2 &: \text{Precio de cierre diario de Multiva,} \\ X_3 &: \text{Precio de cierre diario de Banorte,} \\ X_4 &: \text{Precio de cierre diario de Inbursa,} \\ X_5 &: \text{Precio de cierre diario de Banbajío,} \\ X_6 &: \text{Precio de cierre diario de Value y} \\ X_7 &: \text{Precio de cierre diario de Profuturo.} \end{aligned}$$

En efecto, se define el modelo de regresión lineal por mínimos cuadrados:

$$\boldsymbol{Y} = \beta_0 + \beta_1 \boldsymbol{X}_1 + \beta_2 \boldsymbol{X}_2 + \beta_3 \boldsymbol{X}_3 + \beta_4 \boldsymbol{X}_4 + \beta_5 \boldsymbol{X}_5 + \beta_6 \boldsymbol{X}_6 + \beta_7 \boldsymbol{X}_7,$$

donde $\boldsymbol{X_k} = [1, x_{1,k}, x_{2,k}, ..., x_{692,k}]^t$, para toda $k = 1, \ldots, 7$, $\boldsymbol{Y} = [y_1, y_2, \ldots, y_{692}]^t$.

A través del paquete lm de *R*, se computan los parámetros β_k y sus respectivos niveles de significancia p-valor.

En el Cuadro 5.1 se visualizan los coeficientes β_k y sus respectivos niveles de significancia para la pruebas de hipótesis a un nivel de significancia $\alpha = 0.05$. Se tiene que para las emisoras Inbursa (X_4) y Value (X_6), los precios de cierre tienen estadísticamente poca o nula relación con el precio de cierre de Santander (Y). Por el contrario, el resto de las emisoras, estadísticamente se puede inferir que sus precios de cierre afectan el precio de Y.

Emisora	Predictor/Métrica	Coeficiente	Significancia
-	*Intercepto*	133.56	<2e-16
BBVA	X_1	-4.91	0.0000000000000994
Multiva	X_2	-0.19	0.000000000509
Banorte	X_3	0.95	<2e-16
Inbursa	X_4	0.38	0.175644
Banbajío	X_5	0.58	0.000256
Value	X_6	-0.04	0.08317
Profuturo	X_7	-0.67	<2e-16

Cuadro 5.1: Resultados del primer modelo de regresión lineal. Fuente: Elaboración propia. Yahoo! Finance.

Por este motivo, se ajusta un nuevo modelo, excluyendo aquellas emisoras cuyo p-valor<0,05.

En efecto, se define un nuevo modelo donde:

$$
\begin{aligned}
Y &: \text{Precio de cierre diario de Santander,} \\
X_1 &: \text{Precio de cierre diario de Multiva,} \\
X_2 &: \text{Precio de cierre diario de BBVA,} \\
X_3 &: \text{Precio de cierre diario de Banorte,} \\
X_4 &: \text{Precio de cierre diario de Banbajío,} \\
X_5 &: \text{Precio de cierre diario de Profuturo.}
\end{aligned}
$$

En el Cuadro 5.2, se tienen un conjunto de predictores para el modelo de regresión cuyos coeficientes β_k son significativos para la prueba de hipótesis de no nulidad (ver Sección 4.3.2). Por lo que se pueden hacer los siguientes comentarios acerca de los valores de los coeficientes:

Emisora	Predictor/Métrica	Coeficiente	Significancia
-	*Intercepto*	142.35	<2e-16
Multiva	X_1	-5.56	2e-16
BBVA	X_2	-0.18	3.64e-13
Banorte	X_3	0.95	<2e-16
Banbajío	X_4	0.72	1.98e-09
Profuturo	X_5	-0.69	2e-16

Cuadro 5.2: Resultados del segundo modelo de regresión lineal. Fuente: Elaboración propia con datos de Yahoo! Finance.

En cuanto a Banorte y Banbajío, sus valores son positivos (0.95 y 0.72, respectivamente). Esto implica que los valores de estas emisoras afectan de manera positiva al precio de cierre de Santander. Puesto que para Banorte, si el precio de cierre de esta emisora aumenta en una unidad, el precio de Santander aumenta en 0.95 unidades. Análogamente, si Banbajío aumenta en una unidad su valor, Santander aumentará en 0.72 unidades su precio.

El comentario anterior es análogo para el caso en el que los coeficientes son negativos (Multiva, BBVA y Profuturo). Particularmente, cabe señalar que para el caso de Multiva, un cambio unitario en el precio de cierre de esta emisora, causa un decremento en Santander en 5.56 unidades.

Con los resultados mostrados en los Cuadros 5.1 y 5.2, se hizo una discriminación de variables tomando únicamente emisoras cuya significancia exceda 0.05. Esto se realiza con el motivo de conocer predictores que realmente tengan un efecto significativo sobre el cambio de precio en las acciones de Santander.

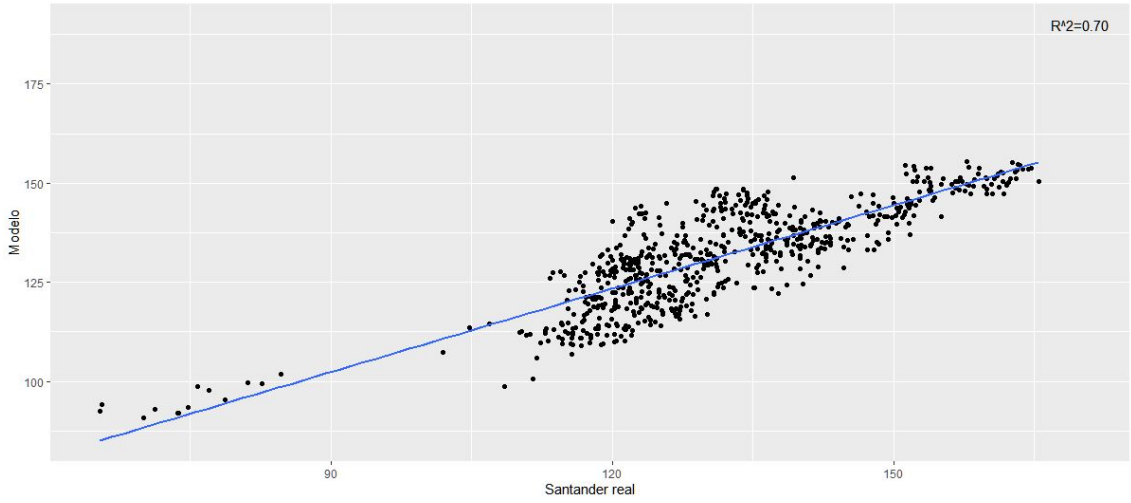

Figura 5.1: Diagrama de dispersión entre valores observados y estimados por el segundo modelo de regresión. Fuente: Elaboración propia con precios de Yahoo! Finance

En la Figura 5.1 se muestra la gráfica de dispersión de los precios de cierre de Santander (valores observados) en el eje x. En el eje y se visualiza el valor estimado por el segundo modelo de regresión. De esta visualización se puede inferir que de tal comparación, se obtiene aproximadamente la recta $y = x$, con una $R^2 = 0.7$, lo cual indica que el modelo de regresión explica el 70 % de la varianza de la variable dependiente y (precios de cierre de Santander). Por lo tanto, es posible inferir que el precio de cierre de las acciones del sector analizado, explican en gran medida los precios de cierre de Santander.

En otras palabras, como se había supuesto, un parámetro a tomar en cuenta cuando se sigue el precio de una acción en particular es seguir precios de acciones similares. En este caso, del mismo sector.

5.2. Aplicación del modelo de regresión logística

Una vez aplicado el análisis exploratorio mostrado en la Sección 4.2, se ha logrado obtener una base con variables significativas ante el precio de Santander. Ahora, se aplicará el modelo de regresión logística para predecir la variación futura de las acciones.

5.2.1. Construcción de variables independientes

La idea es explicar la dirección del precio de la acción de Santander, en función del movimiento del mercado durante dos días hábiles anteriores. Por lo que tal movimiento viene descrito como las ganancias de las otras emisoras del sector bancario en tal ventana temporal. Dicho en palabras llanas, se pretende observar las ganancias de las emisoras del sector bancario durante un periodo $t-h$, y predecir la dirección del precio en un momento t. Donde t y h son días hábiles. $t \leq 2$

Sean:

- $A_{i,j}$: Precio de apertura para $i = 1, ..., n$ y la $j-$ésima emisora.
- $C_{i,j}$: Precio de cierre para $i = 1, ..., n$ y la $j-$ésima emisora.

Se define la i-ésima ganancia o pérdida de la j-ésima emisora en al día i como $X_{i,j} = C_{i,j} - A_{i,j}$, es decir, las ganancias diarias de las emisoras Multiva, BBVA, Banorte, Banbajío y Profuturo. Asimismo, $D_i = C_{i,j} - A_{i,j}$ es la ganancia o pérdida de Santander en la observación i.

5.2.2. Construcción de variable dependiente

Por la construcción del modelo de regresión logística marcado en la Sección 4.3.3, se define la variable independiente binaria Y que representa si la acción subió o bajó de precio en un día. Por lo tanto,

$$Y_i : \begin{cases} 0, & \text{si } D_i < 0 \\ 1, & \text{si } D_i \geq 0 \end{cases}$$

5.2.3. Construcción de parámetros del modelo

Puesto que se busca predecir el movimiento de la acción de Santander (sube o baja), basado en el comportamiento del mercado, dos y un día anteriores a la observación i, se definen los vectores de observaciones:

Emisora	Vector de Observaciones
Multiva	X_1
BBVA	X_2
Banorte	X_3
Banbajío	X_4
Profuturo	X_5
Santander	X_5

Cuadro 5.3: Vectores de observaciones.

Se ajusta el modelo

$$P(Y|X_1, ..., X_5) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + ... \beta_6 X_6)}}.$$

Por otro lado, vale la pena mencionar que los vectores $\boldsymbol{X}_k$, $k \leq 5$, son observaciones correspondientes al momento t, por lo que se desfasan las observaciones uno y dos días, de tal manera que el modelo al tiempo t es de la forma:

$$P(Y_t|X) = \frac{1}{1 + e^{-(\beta_1 X_{1,t-1} + \beta_2 X_{1,t-2} + ... + \beta_1 1 X_{6,t-1} + \beta_{12} X_{6,t-2})}}.$$

Se estiman 12 parámetros, puesto que son 6 acciones que conforman al portafolio, multiplicado por 2, dado que se ocupan las observaciones desfasadas 1 y 2 días.

Después de limpieza y consolidación del conjunto de datos, el modelo se queda con $n = 660$ observaciones, correspondientes al periodo del 29 de septiembre de 2017 al 1 de junio de 2020.

Para evaluar la precisión del modelo, se separa la muestra en dos: conjunto de entrenamiento y conjunto de validación. Se toma una muestra aleatoria del 80 % de la base total (80 % para entrenamiento y 20 % para validación).

El proceso señalado se realiza con el motivo de ajustar los parámetros del modelo de regresión (entrenamiento) con una muestra representativa del 80 % de toda la base y el remanente 20 % para medir calidad y precisión de las predicciones mencionadas en la Sección 4.4.

En otras palabras, se ajustan los parámetros del modelo con observaciones correspondientes al 29 de septiembre de 2017 al 25 de octubre de 2019. Los datos remanentes que corresponden al 28 de octubre de 2019 al 1 de junio de 2020, se utilizan como conjunto de prueba (métricas de precisión y validación.).

Esto es, se busca calcular la probabilidad de que las acciones de Santander respondan favorablemente o desfavorablemente, dado el comportamiento del sector bancario de dos días hábiles anteriores. Y posteriormente se pone a prueba el modelo con el conjunto de prueba, cuyas direcciones del precio de Santander se saben de antemano.

En efecto, con la aplicación del método *LogisticRegression* del paquete *scikitlearn* de *Python*, se obtienen los resultados mostrados en el Cuadro 5.4.

Emisora	Predictor/Métrica	Coeficiente
Multiva	$X_{1,t-1}$	-0.0432
Multiva	$X_{1,t-2}$	0.3716
BBVA	$X_{2,t-1}$	0.0817
BBVA	$X_{2,t-2}$	-0.0382
Banorte	$X_{3,t-1}$	0.0423
Banorte	$X_{3,t-2}$	-0.0041
Banbajío	$X_{4,t-1}$	-0.2242
Banbajío	$X_{4,t-2}$	0.0603
Profuturo	$X_{5,t-1}$	0.0538
Profuturo	$X_{5,t-2}$	0.0022
Santander	$X_{6,t-1}$	0.1319
Santander	$X_{6,t-2}$	0.1113

Cuadro 5.4: Resultados del modelo de regresión logística. Fuente: Elaboración propia con datos de Yahoo! Finance.

La interpretación del valor de los parámetros $\boldsymbol{\beta}$ consiste en observar el modelo de regresión, pues:

$$\ln\left(\frac{p}{1-p}\right) = \beta_1 X_{1,t-1} + \beta_2 X_{1,t-2} + ... + \beta_{11} X_{6,t-1} + \beta_{12} X_{6,t-2}.$$

Entonces, si $f(x_1, ..., x_{12}) = ln(\frac{p}{1-p})$, se tiene que si se asume que los predictores son constantes, esto lleva a que:

$$\frac{\partial f}{\partial x_i} = \beta_i,$$

para $i \leq 12$. Eso significa que un cambio en el predictor x_i incrementa o disminuye en β_i unidades las *log-odds* de que $Y = 1$. En términos del contexto de la investigación, esto es cuánto varían las

log-odds de que la acción de Santander suba de precio, dada la evidencia que proveen las observaciones x_i.

Retomando los resultados en el Cuadro 5.4 se tiene que el aumento de la ganancias observadas en las emisoras Multiva ($X_{1,t-1}$), BBVA ($X_{2,t-2}$), Banorte ($X_{3,t-2}$) y Banbajío ($X_{4,t-1}$) tienen un efecto adverso a las probabilidades de que las acciones de Santander suban de precio. Análogamente el modelo sugiere que estadísticamente las probabilidades de que Santander suba de precio incrementan cuando Multiva ($X_{1,t-2}$), BBVA ($X_{2,t-1}$), Banorte ($X_{3,t-1}$), Banbajío ($X_{4,t-2}$), Profuturo ($X_{5,t-1}$), Profuturo ($X_{5,t-2}$), Santander ($X_{6,t-1}$) y Santander ($X_{6,t-2}$). Siendo estas últimas 2 algo esperado pues es natural que la acción suba de precio si en el corto plazo, ha mantenido esa tendencia.

Los resultados obtenidos del modelo de regresión logística pueden ser aplicados como sigue: si se sabe que en algún momento las ganancias observadas de $x_{1,t-1}$, $x_{1,t-2}$, $x_{2,t-1}$, $x_{2,t-2}$, $x_{3,t-1}$, $x_{3,t-2}$, $x_{4,t-1}$, $x_{4,t-2}$, $x_{5,t-1}$, $x_{5,t-2}$, $x_{6,t-1}$ y $x_{6,t-2}$ son de \$0.2, \$2.0, \$0.5, \$0.6, \$1.0, \$0.8, \$0.9, \$1.2, \$0.7, \$0.5, \$2.0, \$2.0, respectivamente, entonces la probabilidad de que la acción de Santander suba, dada la evidencia observada es, sustituyendo valores de entrada y parámetros en el modelo de regresión logística:

$P(Y_t|X) = 0.766$.

Esto es, dados los valores de entrada propuestos, existe una probabilidad de que la acción de Santander suba con probabilidad 0.77 el día hábil inmediato siguiente.

Pero antes de dar por hecho que el modelo arroja predicciones precisas, es necesario aplicar el modelo ROC y su correspondiente métrica AUC.

5.3. Aplicación de métricas de precisión

Las predicciones hechas por el modelo de regresión logística están dadas en términos de una probabilidad que naturalmente toma valores en $[0, 1]$. Es por esto que tal probabilidad también puede ser vista como un punto de corte para un modelo de clasificación en un modelo de aprendizaje supervisado. Como se ha comentado con anterioridad, la métricas de precisión son evaluadas en el conjunto de prueba (20 % de la base), dados los parámetros ajustados con el modelo de entrenamiento (80 %).

El punto de corte en este sentido es el valor en el cual, se debería dar una probabilidad de éxito como verdadera. Esto es, para qué valor entre 0 y 1, el modelo es capaz de maximizar la tasa de efectividad del modelo.

Una vez seleccionado este punto de corte, se debe evaluar la especificidad y sensibilidad del modelo a través de la curva ROC y su área bajo la curva (AUC) definidos en la Sección 4.4.2 y 4.4.3. Tanto la traza de la curva ROC, como su área bajo la curva (AUC), fueron calculadas a través del paquete *scikitlearn* del lenguaje libre *Python*. Los códigos de las métricas de validación son mostrados al final de este documento, en el Anexo 7.2.

En la Figura 5.2 se muestra la tasa de predicción del modelo en el eje Y, dado un punto de corte de probabilidad en el eje X.

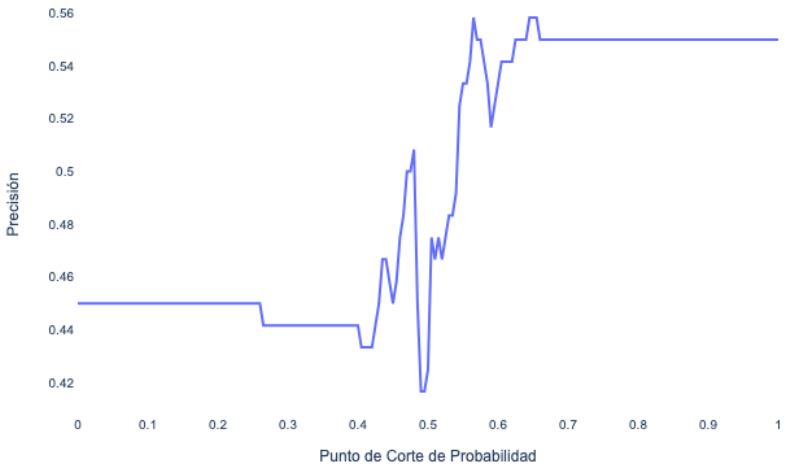

Figura 5.2: Gráfica de precisión del modelo de regresión. Fuente: Elaboración propia con datos de Yahoo! Finance.

Se tiene que para un punto de corte con valor de 0.55, se tiene una tasa de precisión de 0.58 (58 %). Este punto de corte deja la matriz de cofusión en la Ecuación 5.3.1.

		Predicho	
		P	N
Observado	P	36	30
	N	21	33

(5.3.1)

Como se explicó en la Sección 4.4.1, las cantidades de Verdaderos Positivos (VP), Falsos Positivos (FP), Falsos Negativos (FN) y Veraderos Negativos (VN). Esto, se tienen los siguientes valores:

$$\begin{aligned} VP &= 36, \\ FN &= 30, \\ FP &= 21, \\ VN &= 33, \end{aligned}$$

y con ellos se puede calcular la precisión, sensitividad y especificidad del modelo como se muestra a continuación:

$$\begin{aligned} P &= (36+33)/(36+30+21+33) = 0.58, \\ S &= 36/(36+30) = 0.55, \\ E &= 33/(33+21) = 0.61. \end{aligned}$$

Las métricas mostradas indican que el modelo ha hecho una predicción adecuada (aunque mejorable) sobre el conjunto de datos de validación, pues la precisión total del modelo (P) arroja un valor superior

a 50 % y la sensitividad (S) y especificidad (E) son valores superiores al 50 % y no difieren considerablemente una de la otra. Este último hecho es importante de mencionar puesto que si se hubiese escogido un punto de corte superior a 0.55, el modelo hubiera clasificado una mejor sensitividad, pero sacrificando una gran cantidad de especificidad, pues en otras palabras se está siendo excesivamente exigente en cuanto a la cantidad de verdaderos positivos que se desean obtener.

Finalmente, se evaluará la calidad del modelo de aprendizaje mediante las métricas que miden el rendimiento del modelo. Esto es, se calcularán la característica de funcionamiento del receptor (ROC) y el área bajo la curva, ver Sección 4.4.2 y Sección 4.4.3. Para esto se utilizará el paquete *scikitlearn* en *Python*.

Nuevamente el apoyo de las representaciones gráficas permiten decidir si existen un punto en que el compromiso entre ambas métricas es satisfactorio.

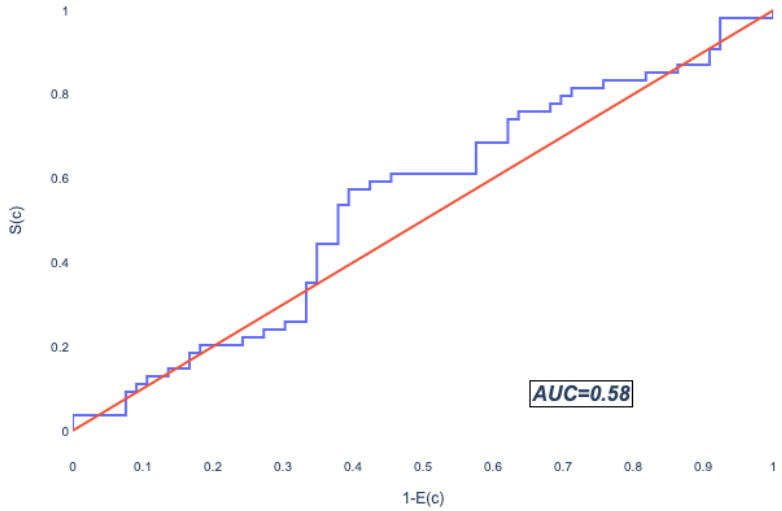

Figura 5.3: Curva ROC del modelo de regresión logística. Fuente: Elaboración propia con datos de Yahoo! Finance.

La Figura 5.3 muestra que para cualquier antiespecificidad y cualquier sensitividad, el modelo tiene un mejor desempeño que la clasificación o predicción al azar (echar un volado). Por otro lado, al calcular su área bajo la curva se obtiene un valor de 0.58.

Con estos dos argumentos y las métricas mostradas se puede inferir que el modelo a pesar de no tener un desempeño excelente (precisión y área bajo la curva mayores o iguales al 80 %), se tiene evidencia suficiente para sugerir que la consideración de variables adicionales al movimiento del mercado podrían hacer que el modelo sea mejor entrenado y por lo tanto mostrar un mejor desempeño.

Capítulo 6

Discusión

Los modelos de clasificación binaria mencionados en la Sección 3.3.1, se han ocupado ampliamente en el análisis bursátil y la econometría. Particularmente (Syed et al, 2018), utiliza un modelo de regresión logística con el propósito de evaluar la calidad de una acción sujeta a diversas cifras financieras. Los resultados muestran que este modelo con una precisión del 88.7 % puede ser utilizado por interesados en el mercado busátil para mejorar sus predicciones.

El presente trabajo desarrolló un modelo de regresión logística, con un nivel de precisión del 58 % que claramente es inferior a lo mostrado por (Syed et al, 2018). Sin embargo, en la Sección 5.3, se amplió el análisis hacia las métricas de precisión, especificidad, sensibilidad, ROC y AUC en una muestra independiente a la muestra utilizada para el ajuste de parámetros del modelo (entrenamiento).

Por lo que se debe comentar que no necesariamente un modelo con buena métrica de precisión implica predicciones correctas en el largo plazo, pues la métrica AUC sirve como una métrica más certera para evaluar la calidad del modelo ya que toma en cuenta tanto a los falsos positivos, como a los falsos negativos.

La construcción del modelo de regresión logística de esta investigación, tomó a la variaciones recientes (2 días) de precio de acciones del mismo sector a la de Santander México. Es decir, esta investigación asumió desde un principio que estas variables de entrada influirían en la variación de Santander. Otra forma de abordar el problema es el de recolectar datos de distintas fuentes y realizar una selección de variables significativas.

Hakob (2018), realizó la predicción de la tendencia de una acción a través de una base de variables que incluye precio de apertura, cierre, bandas de Bollinger, media móvil, entre otros, poniendo a la tendencia de la acción como variable dependiente. El autor utilizó como discriminador de variables (selección de variables o en inglés *feature selection*), el coeficiente de correlación entre cada una de las variables independientes con la variable dependiente. En la investigación presente se utilizó la significancia de los predictores de un modelo de regresión lineal para discriminar variables significativas.

En resumen, el presente trabajo puede ser ampliado a integrar más variables independientes y aplicar un *feature selection* apropiado para entrenar un modelo de aprendizaje de máquina supervisado para evaluar la variación de la acción en función de la variación de los predictores.

Con los resultados mostrados y con la comparación de los modelos de (Syed et al, 2018) y (Hakob ,2018), puede aseverarse que la construcción de un modelo para predecir la variación de una acción debe probar varios modelos de clasificación binaria, un proceso de selección de variables y por último, aplicar métricas de precisión y validación para justamente, comparar la calidad de predicciones entre modelos.

Capítulo 7

Conclusión

En los Capítulos 1 y 2 se presentó un caso muy particular del mercado bursátil en México que es el del estudio de la variación del precio de una acción a través del análisis de emisoras del mismo sector. Se partió de la hipótesis de que acciones similares, siguen el comportamiento de accciones similares (en este caso, del mismo sector). La estrategia usada fue utilizar modelos de regresión para medir qué tanto influye el precio de acciones de Santander en función de la variación de precios del mismo sector financiero.

Las pruebas estadísticas en el modelo de regresión lineal mostraron que efectivamente existen algunas emisoras que siguen una tendencia que afecta la cotización de las acciones de Santander. Y lo que es más importante, también se mostró que inclusive el aumento del precio de acciones de Multiva, BBVA y Profuturo tienen un efecto negativo en las acciones de Santander. Esto es, cuando éstas suben de precio, las acciones de Santander tienden a disminuir su cotización (Cuadro 5.2).

Los resultados empleados posteriores al análisis de regresión lineal con la aplicación del modelo de regresión logística mostraron que efectivamente es posible predecir si una acción de Santander subirá de precio con base en el cambio en precio de cotización de acciones del sector financiero. Es decir, se cumplió el objetivo general de la tesis.

La curva ROC y la métrica AUC calculadas y desarrolladas en la Sección 5.3 mostraron resultados favorables (aunque no excelentes) sobre el modelo. Esto puesto que se logró predecir la variación de Santander con datos que no fueron incluidos en el entrenamiento del modelo y el modelo mostró una AUC de 0.58. *i.e.*, existe una probabilidad de 0.58 de que el modelo clasifique correctamente la si la acción de Santander tendrá un cierre favorable, dada una variación $\boldsymbol{X}$ de entrada.

El objetivo general de la tesis fue alcanzado a través de cumplir los objetivos generales de esta investigación, pues a lo largo de los objetivos se logró realizar un estado del arte apropiado como marco teórico y se presentó el contexto de Santander México. Además una parte muy importante para la investigación aquí presente fueron las visualizaciones de datos con gráficas, pues esto dio pie a tener un mejor entendimiento del problema y se logró tener de manera visual cómo se comportan las acciones del sector financiero y cómo se relacionan entre sí. A través de la aplicación de modelos de regresión lineal y logística, visualizaciones y las métricas de precisión y validación se comprobó que efectivamente es posible hacer inferencias sobre el futuro precio de una acción a partir del estudio de acciones en el mismo sector.

Por otro lado, se llegó a aplicar el modelo de regresión logística con motivo de entrenar un modelo que fuera capaz de predecir (o clasificar como positivo) la respuesta de una acción basado en las ganancias del mercado. También, a lo largo de la investigación se mostró que los modelos de aprendizaje de

máquina y las métricas de precisión y validación pueden ser aplicadas en la predicción sobre acciones de la Bolsa Mexicana de Valores.

Se desarrolló un esfuerzo por aplicar un modelo de aprendizaje de máquina en las finanzas bursátiles. Sin embargo, este modelo de regresión logística no es el único que puede ser utilizado para resolver un problema de esta índole. Existen otros modelos como bosques aleatorios, máquinas de soporte vectorial, redes neuronales, entre otros. Se escogió este modelo por la interpretación de sus parámetros y por la sencillez de su aplicación con la librería *stats* de *R*.

Un modelo con mejor validez para una aplicación en la realidad (puesta en producción) debería incluir información en tiempo real, noticias, análisis en redes sociales, etc., para tener mayor la información disponible que ayude a entrenar un modelo que genere mejores predicciones.

7.1. Comentario final

Toda persona y organización se encuentra con el difícil reto de tomar decisiones todos los días y a todas las horas con la premisa de minimizar las pérdidas y maximizar las ganancias asociadas a esta decisión. El mercado de valores no es ajeno a esta constante, puesto que cualquier tenedor de una acción se ve en la necesidad de vender o comprar a un precio que traiga ganancias positivas.

En este sentido, el aprendizaje de máquina ha resultado en un conjunto de herramientas muy útiles para con base en evidencia estadística disponible, se buscan hacer inferencias hacia el futuro, tratando de utilizar un método que se base en procedimimientos lo más objetivos posibles. Esto pues, con la intención de tomar decisiones basadas en evidencia e inferencia científica y no solamente con conocimiento empírico y/o subjetivo.

Sin embargo, estos esfuerzos por desarrollar modelos de aprendizaje de máquina son poco eficientes si no se cuenta con una estrategia y propósitos bien definidos, antes de aplicar un modelo de regresión o clasificación.

Siendo el mercado de valores de naturaleza completamente azarosa, pues no existe manera de saber a ciencia cierta cuál será el precio futuro de una acción. Pues como aseveró (Weatherhall, 2013), uno puede imaginarse al mercado como un juego de azar gigantesco.

Por definición, un fenómeno aleatorio no se puede predecir. Sin embargo, es posible realizar esfuerzos por aprender sobre evidencia disponible para hacer estimaciones en un futuro.

El aprendizaje de máquina y la estadística computacional han introducido modelos ajustables a los diversos problemas que se tienen en la economía, finanzas, etc. El trabajo aquí presentado mostró que a través de un modelo de clasificación binaria (regresión logística), inferencias y estimaciones sobre el futuro se pueden realizar, teniendo una estrategia de negocio y aplicación de los modelos adecuados.

7.2. Otras aplicaciones

Naturalmente, el análisis bursátil no es el único campo de estudio en el cual se puede aplicar el aprendizaje de máquina para obtener predicciones de cierta índole. De hecho, prácticamente estas herramientas han tenido una aplicación en varias ramas de la ciencia.

Noor y Narwal (2017) aplicaron modelos de aprendizaje supervisado y no supervisado para entrenar un modelo de redes neuronales con la finalidad de indicar si una persona tiene cáncer, basado en una serie de variables independientes. Lo cual implica un aporte valioso para las ciencias médicas, pues es posible indicar a tiempo si una persona es propensa a padecer la enfermedad.

En la rama de la informática, se han aplicado los modelos de SVM y Bayes Ingenuo con el fin de predecir si una serie de correos son potencial *spam* (Trivedi, 2016). Lo cual funge como una herramienta útil para cualquier persona u organización que podría sufrir fraude o pérdida debido a malas prácticas con el correo electrónico.

El desarrollo y despliegue de modelos de aprendizaje de máquina ha sido recientemente de alto impacto en diversas disciplinas. Uno de ellos ha sido la aplicación en áreas de negocios y económico administrativas. Mientras las necesidades de los consumidores van cambiando a través del tiempo, la recolección y análisis de datos han sido de gran apoyo para grandes compañías, pues a través de estos análisis se tiene un mejor soporte para la toma de decisiones basada en criterios objetivos como los modelos matemáticos y estadísticos (van Liebergen, 2017).

Finalmente se debe mencionar que el éxito de la aplicación de modelos de aprendizaje de máquina depende naturalmente de aspectos técnicos como la consolidación de las bases de datos de entrena-

miento, el entrenamiento del modelo y la validación del modelo pero también recae en la estrategia y el asesoramiento de expertos en el tema.

Si no se cuenta con un marco teórico que acompañe al modelo, se puede llegar a cometer errores de interpretación que eventualmente llevan a conclusiones erróneas. Pues de acuerdo con el gran estadístico George Box, todos los modelos son erróneos, pero algunos son útiles. Lo cual no significa literalmente que los modelos son inservibles, sino que después de una serie de validaciones, éstos pueden ser puestos en práctica (Box, 1976).

Anexos

Código de regresión lineal

```
setwd('C:/Users/DELL/Dropbox/TESIS')
library(dplyr)
library(ggplot2)
library(magrittr)
library(reshape2)
library(gridExtra)

precios <- readxl::read_excel('DATA/Master.xlsx', sheet = 9)

precios <- precios %>% filter(VALID=='SI')

#Seleccionar variables numricas
precios_modelo <- precios %>% select_if(is.numeric)

# Aplicar modelo de regresin
model <- lm(SANTANDER~MULTIVA + BBVA + BANORTE+BAJIO+PROFUTURO,data = precios_modelo)

fitted_real = data.frame(precios$SANTANDER,model$fitted.values)

names(fitted_real) = c('Santander real','Modelo')

ggplot(fitted_real,aes(x=`Santander real`,y=Modelo)) + geom_point()+
  geom_smooth(se = F,formula = y~x,method = 'lm') +
  annotate('text',x=170,y=190,label='R^2=0.70')

cor(fitted_real)^2
```

Código de extracción de precios

```
import yfinance as yf
import pandas as pd
import numpy as np
from pandas.tseries.offsets import BDay
from sklearn.linear_model import LogisticRegression
from sklearn import metrics
import matplotlib.pyplot as plt
import seaborn as sns
sns.set()
import plotly.graph_objects as go

portafolio = []
acciones = ['GFMULTIO.MX','BBVA.MX','GFNORTEO.MX','BBAJIOO.MX','GPROFUT.MX','BSMXB.MX',]

for accion in acciones:
    temp = yf.download(accion,
                    start='2015-06-04',
                    end='2020-10-01',
                    progress=False)
    temp.reset_index(inplace=True)
    temp['Date+1'] = temp.Date.apply(lambda x: x + BDay(1))
    temp['Date+2'] = temp.Date.apply(lambda x: x + BDay(2))
    temp[f'Ganancia {accion}'] = temp['Close'] - temp['Open']
    portafolio.append(temp)

santander_pred,santander = portafolio[5].copy(),portafolio[5].copy()

santander_pred.dropna(inplace=True)

santander_pred['Y'] = np.where(santander_pred['Close'] - santander_pred['Open'] > 0,1,0)

santander_pred = santander_pred[['Date','Y']]

for count, data in enumerate(portafolio):
    santander_pred = santander_pred.merge(data[['Date+1',f'Ganancia
        {acciones[count]}']].rename(columns={'Date+1':'Date'}),how='left',on='Date')
    santander_pred = santander_pred.merge(data[['Date+2',f'Ganancia
        {acciones[count]}']].rename(columns={'Date+2':'Date'}),how='left',on='Date',suffixes=("-1",
        "-2"))
santander_pred.dropna(inplace=True)
```

Aplicación de modelo de regresión logística

```
import yfinance as yf
import pandas as pd
import numpy as np
from pandas.tseries.offsets import BDay
from sklearn.linear_model import LogisticRegression
from sklearn import metrics
import matplotlib.pyplot as plt
import seaborn as sns
sns.set()
import plotly.graph_objects as go

tam = len(santander_pred)
s_train = round(0.8*tam)
s_test = tam - s_train

print(tam,s_train,s_test)

df_train = santander_pred.iloc[:s_train]
df_test = santander_pred.iloc[s_train:s_train+s_test]

print(len(df_train),len(df_test))

X_train = df_train.drop(['Date','Y'],axis=1).values
y_train = df_train['Y'].values
X_test = df_test.drop(['Date','Y'],axis=1).values
y_test = df_test['Y'].values

logreg = LogisticRegression(fit_intercept=False)
logreg.fit(X_train, y_train)
```

Códigos de métricas de validación

```
import plotly.express as px
import kaleido
import orca

fig = px.line(x=th, y=acc, title='Precisin')
fig.update_layout(
    xaxis_nticks=20,
    title="",
    xaxis_title="Punto de Corte de Probabilidad",
    yaxis_title="Precisin",
    legend_title="",
    font=dict(
        family="Arial",
        size=10,
    ),
    paper_bgcolor = 'rgba(255,255,255,1)',
    plot_bgcolor = 'rgba(255,255,255,1)'
)

fig.show()
fig.write_image("../Plots/precision_plot.png")

from sklearn.metrics import confusion_matrix
confusion_matrix = confusion_matrix(y_test, y_pred)
print(confusion_matrix)

from sklearn.metrics import classification_report
print(classification_report(y_test, y_pred))

from sklearn.metrics import roc_auc_score
from sklearn.metrics import roc_curve
logit_roc_auc = roc_auc_score(y_test, logreg.predict(X_test))
fpr, tpr, thresholds = roc_curve(y_test, logreg.predict_proba(X_test)[:,1])
```

Códigos de gráficas

```
setwd('C:/Users/DELL/Dropbox/TESIS')
library(dplyr)
library(ggplot2)
library(magrittr)
library(reshape2)
library(gridExtra)

precios <- readxl::read_excel('DATA/Master.xlsx', sheet = 9)

precios <- precios %>% filter(VALID=='SI')


melt_precios <- melt(data = precios,
id.vars= "Fecha",
measure.vars = c("SANTANDER", "MULTIVA", "BBVA", "BANORTE", "INBURSA", "BAJIO", "VALUE",
    "PROFUTURO"
))

melt_precios$value %<>% sapply(as.numeric)

# Facet plot precio ------------------------------------------------------
melt_precios %>%
  ggplot(aes(x=value))+
  geom_density(fill='turquoise')+ggtitle("")+xlab(paste("Precio de cierre"))+
  ylab("Densidad")+labs(fill='') + facet_wrap(~variable, ncol=2)+
  theme(strip.text = element_text(size=10))


# Corr plot --------------------------------------------------------------

precios %>% select_if(is.numeric) %>%
  cor() %>% round(2) %>% melt() %>%
  ggplot(aes(x=Var1, y=Var2, fill=value)) +
  geom_tile(color='white')+
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                       midpoint = 0, limit = c(-1,1), space = "Lab",
                       name="Correlacin de Pearson") +
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 45, vjust = 1,
                                   size = 12, hjust = 1))+
  coord_fixed()+xlab('Emisora')+ylab('Emisora')


# Violin plot ------------------------------------------------------------

melt_precios %>% ggplot(aes(x=variable,y=value))+
  geom_violin() + stat_summary(fun.y=median, geom="point", size=2, color="red")+
  xlab('Variable') + ylab('Precio de cierre')
```

Bibliografía

Altman E.I. (1968). *Financial ratios, discriminant analysis and the prediction of corporate.*

Banco de México.(2020). *Mercados Financieros* Banco de México, recuperado de http://educa.banxico.org.mx/banco_mexico_banca_central/sist-finc-mercados-financiero.html

Bartels R. (2005). *Re-interpreting R^2, regression through the origin, and weighted least squares*, University of Sydney Business School, pp. 6-10.

Banco Bilbao Viscaya (2020). *Instrumentos Financieros* Banco Bilbao Viscaya, recuperado de https://www.bbva.com/es/instrumentos-financieros-todos/

Bolsa Institucional de Valores (2020). *Acerca de BIVA* Bolsa Intsitucional de Valores, ecuperado de https://www.biva.mx/nosotros/acerca_de

Bolsa Mexicana de Valores (2019). *El Índice de Precios y Cotizaciones y su importancia para el mercado* Bolsa Mexicana de valores, recuperado de https://blog.bmv.com.mx/2019/03/21/el-indice-de-precios-y-cotizaciones/.

Bolsa Mexicana de Valores (2020). *Clasificación* Bolsa Mexicana de valores, recuperado de https://www.bmv.com.mx/es/mercados/clasificacion.

Bodie Z., Kane A., Marcus A.(2011). *Investments* McGraw-Hill, Nueva York.

Box G.E.P. (1976). *Science and Statistics*, ournal of the American Statistical Association, Vol. 71, No. 356. (Dec., 1976), pp. 791-79. Recuperado de http://www-sop.inria.fr/members/Ian.Jermyn/philosophy/writings/Boxonmaths.pdf

Canelles A. (2017). *Análisis técnico de mercados basado en técnicas de inteligencia artificial*. Universidad de Murcia.

Chen R., Zheng Z. (2011). *Unbiased Estimation, Price Discovery, and Market Efficiency: Futures Prices and Spot Prices* Systems Engineering - Theory Practice, Volume 28, Issue 8, pp. 2-11.

Deakin, E. (1972). A Discriminant Analysis of Predictors of Business Failure. Journal of Accounting Research, 10(1), 167-179. doi:10.2307/2490225

Del Valle A.(2015). *Curvas ROC (Receiver-Op erating-Characteristic) y sus aplicaciones*, Universidad de Sevilla, 2015.

Diario Oficial de la Federación (2008). *Resolución por la que se autoriza la organización y operación de una institución de banca múltiple filial denominada Banco Santander (México), S.A., Institución de Banca Múltiple, Grupo Financiero Santander.*, recuperado de http://dof.gob.mx/nota_detalle.php?codigo=5044611&fecha=11/06/2008

Días O. (2005). *Aplicación y Estudio de los métodos utilizados por el Análisis Técnico y Fundamental para la inversión en acciones* , Universidad de las Américas Puebla.

Dobson A.(2008). *An Introduction of Generalized Linear Models*, Chapman and Hall.

El Economista (2019). *Seis emisoras acaparan a la BMV*, Diario El Economista, recuperado de https://www.eleconomista.com.mx/mercados/Seis-emisoras-acaparan-a-la-BMV-20191003-0104.html.

El Financiero (2018). *En México, 54 millones de personas tienen un producto financiero: ENIF*, Diario El Financiero, recuperado de https://www.eleconomista.com.mx/sectorfinanciero/En-Mexico-54-millones-de-personas-tienen-un-producto-financiero-ENIF-20181125-0044.html.

Forbes (2017). *Los 10 bancos más grandes de México*, Revista Forbes, recuperado de https://www.forbes.com.mx/los-10-bancos-mas-grandes-de-mexico/.

Forbes (2017). *Banco de Londres, México y Sudamérica, el primer banco comercial de México*, Revista Forbes, recuperado de https://www.forbes.com.mx/brand-voice/banco-londres-mexico-sudamerica-primer-banco-comercial-mexico/.

Golberg M. y Cho H. (2010). *Introduction to Regression Analysis* Recuperado de https://www.researchgate.net/publication/264700780_Introduction_to_Regression_Analysis

Hakob G. (2017). *Stock Market Trend Prediction Using Support Vector Machines and Variable Selection Methods*, 10.2991/ammsa-17.2017.45. Recuperado de https://www.researchgate.net/publication/318354741_Stock_Market_Trend_Prediction_Using_Support_Vector_Machines_and_Variable_Selection_Methods

Hiba Satia K., Aditya S., Adarsh P., Sarmista P., Saurav S.(2019). *Stock Market Prediction Using Machine Learning Algorithms* International Journal of Engineering and Advanced Technology (IJEAT)ISSN: 2249 –8958,Volume-8.

Karch J.(2019). *Improving on Adjusted R-Squared* Leiden University. Recuperado de https://www.researchgate.net/publication/335845928_Improving_on_Adjusted_R-Squared. 10.31234/osf.io/v8dz5.

Ladrón de Guevara R. (2004). *El Indice De Precios Y Cotizaciones (Ipc) De La Bolsa Mexicana De Valores: Importancia De Los Indicadores Financieros En Los Mercados Bursatiles* Ciencia Administrativa. 1. 154-175. Recuperado de https://www.researchgate.net/publication/279181242_EL_INDICE_DE_PRECIOS_Y_COTIZACIONES_IPC_DE_LA_BOLSA_MEXICANA_DE_VALORES_IMPORTANCIA_DE_LOS_INDICADORES_FINANCIEROS_EN_LOS_MERCADOS_BURSATILES/citation/download

Liu Q., Wu Y. (2012). *Supervised Learning* 10.1007/978-1-4419-1428-6-451.

Low A. W.(2015). *What Is An Index?* Massachusets Institute of Technology, recuperado de https://alo.mit.edu/wp-content/uploads/2015/10/index_5.pdf, p. 3.

Marsland S. (2015) *Machine Learning. An Algorithmic Perspective.* Chapman and Hall/CRC, Nueva York.

Martínez-Camblor P. (2007),*Comparación de pruebas diagnósticas desde la curva ROC*, Revista Colombiana de Estadística, 30 p. 163-176.

McCullagh P.(2000). *Generalized Linear Models*, Chapman and Hall.

MexDer (2017). *Una Introducción a los Mercados*, Bolsa Mexicana de Valores.

Montgomery D. C., Peck E. A., Vining G. G. (2002). *Introducción al Análisis de Regresión Lineal*, Continental, p. 67.

Nanda S.R., Mahanty B., Tiwari M.K. (2010). *Clustering Indian stock market data for portfolio management* Expert Systems with Applications .

Nasteski V. (2017). *An overview of the supervised machine learning methods* B. 4. 51-62. 10.20544/HORIZONS.B.04.1.17.P05.

Noor M., Narwall V. (2017). *Machine Learning Approaches in Cancer Detection and Diagnosis: Mini Review.*, 10.13140/RG.2.2.27775.51363. Recuperado de https://www.researchgate.net/publication/320947210_Machine_Learning_Approaches_in_Cancer_Detection_and_Diagnosis_Mini_Review

Prado H., Ferneda E., Morais L., Luiz A., Matsura E. (2013). *On the Effectiveness of Candlestick Chart Analysis for the Brazilian Stock Market* Procedia Computer Science . 22. 10.1016/j.procs.2013.09.200.

R Core Team (2020). *R: A language and environment for statistical computing.* http://www.R-project.org/. R Foundation for Statistical Computing, Vienna, Austria.

Rawlings O. (2005) *Applied Rgression Analysis*, Springer.

Restrepo B., González J (2007). *De Pearson a Spearman* Revista Colombiana de Ciencias Pecuarias, vol. 20, núm. 2, abril-junio, 2007, pp. 183-192. Recuperado de https://www.redalyc.org/pdf/2950/295023034010.pdf.

Ross S.A., Westerfield R.W., Jordan B.D (2010). *Fundamentos de Finanzas Corporativas* McGraw-Hill, Nueva York, pp. 438-447.

Ross S., Westerfield R., Jaffre J. (2012). *Finanzas corporativas.* Mc Graw-Hill, Nueva York.

Banco Santander México.(2020). *Glosario Financiero* Santander, recuperado de https://www.santander.com.mx/PDF/canalfin/documentos/glosario.pdf, p. 2.

Santander México (2020). *Historia*, Banco Santander México, recuperado de https://www.santander.com.mx/ir/historia.

Smola A. (2008). *Introduction to Machine Learning*, Cambridge University Press.

Syed S., Mubeen M., Lal A., (2018). *Prediction of stock performance by using logistic regression model: evidence from Pakistan Stock Exchange (PSX)*, Asian Journal of Empirical Research. 8 .10.18488/journal.1007/2018.8.7/1007.7.247.258.

Trivedi S. (2016). *A study of machine learning classifiers for spam detection.*, 176-180. 10.1109/ISCBI.2016.7743279. Recuperado de https://www.researchgate.net/publication/310498804_A_study_of_machine_learning_classifiers_for_spam_detection

van Liebergen B. (2017). *Machine Learning: A Revolution in Risk Management and Compliance?*, Institute of International Finance. Recuperado de https://www.iif.com/portals/0/Files/private/32370132_van_liebergen_-_machine_learning_in_compliance_risk_management.pdf

Velasco R.(2011). *Introducción al Mercado Bursátil*, Universidad de Alicante.

Visa S., Ramsay B., Ralescu A., van Der Knaap E. (2011). *Confusion Matrix-based Feature Selection*, CEUR Workshop Proceedings.

Wackerly D., Mendenhall W., Scheaffer R. (2010). *Estadística Matemática con Aplicaciones* Cengage Learning, pp. 598-600.

Weatherhall D. J.(2008). *Cuando los Físicos Asaltaron los Mercados*, Ariel, pp. 31,32.