Capítulo 2

Marco Teórico

2.1. Economía Colombiana

2.1.1. Introducción

Para la creación de un modelo de predicción de la TRM de Colombia, tomando como hipótesis de trabajo que existe un número finito y reducido de variables de aporte que regulan el valor de la misma a través de los ingresos por exportación y su contribución a la economía nacional, se debe primeramente comprender y definir estos conceptos. La sección del marco teórico que cubre la economía de Colombia, tiene como finalidad abarcar los siguientes temas.

1. Definir correctamente el concepto de tasa de cambio y su específico colombiano, la tasa de mercado representativa, desentrañando la formula que usa la Superintendencia Bancaria para su valuación diaria.

2. Entender las bases del comercio internacional de Colombia y cuales son sus principales productos de exportación, sobre todo con el afán de identificar correctamente candidatos como variables de aporte para alimentar de datos el modelo de aprendizaje automatizado.

3. Por último, especificar el funcionamiento de los elementos financieros derivados de compra de divisas tales como los *forward* y su correspondiente reglamentación bajo la leyes de Colombia.

Entender el funcionamiento de la economía de Colombia, sus principales componentes de exportación, y los marcos legales que rigen las estructuras de la TRM y los productos financieros de compra y venta de divisas, nos da

luces no solo para entender correctamente el problema, sino para plantear propuestas de solución matemáticas que tengan una amplia correlación entre el modelo abstracto y el comportamiento en la vida real del proceso.

2.1.2. La Tasa de Cambio

La moneda de un país tiene una equivalencia en moneda de otro y ese valor se conoce como tasa de cambio. Explicamos el concepto apoyados en los escritos del autor Mauricio Cárdenas [Cárdenas, 2016]. También es importante explicar porqué los productos de exportación tienen un efecto en la canasta de divisas y la balanza de pagos [Carriello, 2010].

2.1.3. La TRM

Dennis Robertson (Robertson, 1922) definió el dinero como *"todo aquello generalmente aceptado para el pago de una obligación"*. El dinero en su forma más simple es el medio de pago de total liquidez, constituido por el *efectivo* (billetes y monedas) y puesto en circulación por la Banca Central y por el *dinero bancario*, correspondiente a los depósitos en bancos comerciales que son transferibles por medio de cheque.

El intercambio de bienes y comercio internacional se realiza tomando como premisa que países con diferente moneda tendrán que llegar a algún tipo de mecanismo para compensar las compras, ventas y pagos de las mismas entre ambos actores. A falta de una moneda común (por lo menos en términos legales) el mecanismo que rige dicha condición de medio de operación es la tasa de cambio.

Concepto: Tasa de Cambio

Para entender el concepto de la tasa representativa de mercado, es importante primero entender el concepto de tasa de cambio. Dicha idea es muy sencilla, y gira entorno al valor de una moneda en relación con otra (Fishwick, Georgiou, Huston, y Marechal, 2010). En épocas pasadas, el tipo de cambio era fijo, pero esta teoría ha quedado atrás con la implementación de tipos de cambio de flotación libre. La preocupación de los gobiernos gira en procurar mantener un determinado tipo de cambio ni estimule la revalorización de la moneda, ni mucho menos genere una devaluación de la misma (Cárdenas, 2016).

Devaluación de la Moneda

Se entiende como devaluación monetaria la pérdida del valor nominal de la moneda nacional frente a otra u otras monedas extranjeras. Las causas generadoras de la devaluación se pueden sintetizar principalmente en dos:

- Falta o disminución de la demanda de la moneda nacional.
- Una mayor demanda de la moneda extranjera por parte de los consumidores y comerciantes de la nación.

En un sistema de cambio libre (dólar de flotación), en el cual la intervención del Banco de la República es nula, la devaluación toma el nombre de *depreciación*.

Apreciación de la Moneda Local

A veces, por causas externas a la economía de un país, la moneda local se ve sobrevaluada, sea por la abundancia de dólares procedentes del exterior o por el ingreso de capitales extranjeros al país. Esto genera que haya más reservas de dólares, provocando que la moneda local se aprecie por la mayor oferta de los capitales extranjeros.

Dólar de Flotación

El 25 de septiembre de 1999, la Junta Directiva del Banco de la República de Colombia optó por desmontar la banda cambiaría y dar paso al dólar flotante. Cuando el precio de la divisa se mueve por libre juego de la oferta y la demanda, sin límite de techos o pisos, se habla de un régimen de flotación. La flotación implica que el Banco de la República no tendrá en adelante ninguna injerencia en la fijación del precio del dólar (Cárdenas, 2016).

Las oportunidades en las cuales el estado ha tratado de varias formas de estabilizar la moneda y el tipo de cambio no han sido pocas. El país ha pasado a través de ciclos de revaluación y devaluación alternados, ambos con impactos negativos para la economía.

1. En el año 2007 el Gobierno Nacional intentó frenar el ingreso de dólares producto de capital golondrina con medidas de cautelas de depósitos de un cuarenta por ciento del valor durante seis meses, tratando de evitar la especulación (Decreto 2466 MINHACIENDA, Junio 2007)

2. La disminución de los capitales y el aumento del desempleo llevó al Gobierno Nacional a desarticular dicha medida en el año 2008 (Decreto 1888 MINHACIENDA, Mayo 2008)

3. En el año 2012 el Banco de la República tomo la estrategia de compras diarias de treinta millones de dólares como forma de mantener la moneda estable y lejos de la apreciación (Empresarios piden bajar tasas de interés por caída del dólar. (Enero 27 del 2012) Portafolio, pp.3)

4. Hacia el año 2014 el Banco de la República, habiendo conseguido su meta de una tasa de cambio estable, redujo notablemente sus esfuerzos de compras de la divisa americana. Lamentablemente hacia mediados del 2015, la caida de los precios del petroleo tuvo un efecto nefasto en la devaluación del peso colombiano, que llegaría a tasas de cambio a finales del año cercanas a los $3,300 pesos.

La TRM (Tasa Representativa del Mercado

La Superintendencia Financiera de Colombia es la que calcula y certifica diariamente la TRM con base en las operaciones registradas el día hábil inmediatamente anterior y la define de la siguiente manera (Circular Reglamentaria Externa del Banco de la República DODM-146, 2015):

> La tasa de cambio representativa del mercado (TRM) es la cantidad de pesos colombianos por un dólar de los Estados Unidos (antes del 27 de noviembre de 1991 la tasa de cambio del mercado colombiano estaba dada por el valor de un certificado de cambio). La TRM se calcula con base en las operaciones de compra y venta de divisas entre intermediarios financie-ros que transan en el mercado cambiario colombiano, con cumplimiento el mismo día cuando se realiza la negociación de las divisas.

La Superintendencia Financiera de Colombia no determina el valor de la TRM sino de un elemento derivado de las operaciones de compra y venta de la misma. Son los agentes de operación (exportadores que venden sus productos en dólares y los deben canjear a pesos colombianos e importadores que compran sus productos en dólares y para tal fin cambian sus pesos colombianos). Ambos obedecen a fuerzas del mercado que dan forma y materializan la valorización.

2.1.4. Exportaciones de Colombia

Si bien no buscamos ser expertos en ninguno de los tipos de exportación que hace Colombia, es importante en esta sección describir uno a uno los rubros con mayor contribución, ya que serán nuestras variables independientes para aplicar en el proceso de aprendizaje automatizado y modelar el comportamiento futuro de la TRM.

El Petróleo

Del petróleo se dice que es el energético más importante en la historia de la humanidad, que es un recurso no renovable que aporta el mayor porcentaje del total de la energía que se consume en el mundo. En cuanto a Colombia, hace parte del grupo de países en el mundo que tiene petróleo, sin llegar a ser un país petrolero; su producción para el año 2015 tan solo alcanzó un millón de barriles diarios, de los cuales no todos son clasificados como los mejores, ya que no alcanzan según las normas API el nivel superior a 26 grados (Cárdenas, 2016).

En Colombia los recursos naturales no renovables, entre ellos, los hidrocarburos, son propiedad del Estado. La política petrolera es definida por el Gobierno Nacional a través del Ministerio de Minas y Energía, y hasta el año 2003 Ecopetrol era la empresa encargada de su ejecución.

El Carbón

Colombia, en cuanto a recursos carboníferos se refiere, ocupa dentro de los países latinoamericanos un lugar privilegiado, pues cuenta con las mayores reservas y cuenta con gran variedad de calidades. Este potencial carbonífero está distribuido en las tres cordilleras principales, correspondiendo la mayor parte a la cordillera oriental [Cárdenas, 2016].

La importancia del carbón colombiano, más que por sus características, es por su posición estratégica (particularmente en las minas de la Guajira), pues facilita el acceso al mercado europeo y norteamericano, y porque ha logrado, con relativo éxito, la conquista de dichos mercados por su precio y calidad respecto al de los carbones procedentes de Australia e Indonesia.

El Café

El café Colombiano es reconocido a nivel mundial a través de su marca registrada Juan Valdez. Dado que es una de las exportaciones que continua

creciendo, es de esperar que sea una fuente de divisas y exista una correlación estrecha entre el precio del café y el valor de la TRM.

El Níquel

La importancia del níquel radica en las aleaciones con otros elementos para dar fuerza y resistencia a la corrosión en amplias variaciones de temperatura. Se utiliza principalmente en aleaciones con el hierro y el acero para las fabricaciones de aceros inoxidables empleados en la industria en forma general. En Colombia, los recursos identificados pertenecen al grupo de las lateritas niquelíferas, producto de la alteración de las rocas ultramáficas del conocido sistema tectónico ofiolítico [Cárdenas, 2016].

2.1.5. Forwards

El siguiente trabajo de investigación no trata sobre opciones de compra de moneda a futuro (conocido como *forwards*). Sin embargo explicamos de forma sucinta qué son y cómo funcionan, ya que el resultado de las predicciones se utilizará muy seguramente para complementar acuerdos de futuros de divisas como medida de control de costos.

2.2. La Ciencia de Datos

La Ciencia de Datos es una disciplina relativamente nueva, inclusive en muchos entornos académicos. El objetivo de este capítulo es el de resumir los aspectos mayores de la ciencia de datos como estudio multidisciplinario cuyo objetivo es el de hacer sentido de la gran cantidad de datos que surgen de nuestro entorno, con miras a modificar los fenómenos del mundo.

2.2.1. Introducción

La ciencia de datos [Zumel and Mount, 2014] utiliza herramientas de otras ciencias empíricas, estadística, análisis matemático, finanzas, técnicas de visualización, inteligencia de negocios, sistemas expertos, aprendizaje automatizado, bases de datos, bioestadística, y ciencia de la computación con la finalidad de procesar y extraer conocimiento de la data, ya sea que esta se encuentre estructurada o no estructurada.

Previo al termino Ciencia de Datos, el matemático John W. Tukey comienza a circular la idea del análisis de datos versus la estadística en su libro *The Future of Data Analysis* (Tukey, J. 1972). La premisa es que la

estadística es una ciencia formal, mientras que el análisis de datos es una ciencia empírica ya que se basa en datos extraídos de la vida real. Tukey sostuvo que de la data debía extraerse hipótesis para evaluación, y que el análisis confirmatorio de datos debia coexistir al lado del análisis exploratorio de datos. Ambos se apoyan en la estadística como disciplina de aplicacion pero estudian objetos diferentes.

La ciencia de datos (Wikipedia, 2016) ha resultado para muchos una disciplina de reciente creación, pero en la realidad este concepto lo utilizó por primera vez el científico danés Peter Naur en la década de los sesenta como sustituto de las ciencias computacionales. En 1974 publicó el libro Concise Survey of Computer Methods 3 donde utiliza ampliamente el concepto ciencia de datos, lo que permitió que se comenzara a utilizar más libremente entre el mundo académico.

Por otro lado, el matemático japones e inventor de la *Metodología de Cuantificación* Chikio Hayashi define sucintamente (Hayashi, C. 1998) la ciencia de datos no solo como un concepto sintético para unificar las disciplinas de la estadística, el análisis de datos, y sus métodos relacionados, sino por la forma en la cual sus resultados se aplican. Esta nueva metodología incluye tres fases: diseño de la data, recolección de la data, y análisis de la misma.

Muchas veces se ha criticado a la ciencia de datos como el uso disimulado de estadística bajo un nombre diferente con fines comerciales. En 2001, William S. Cleveland introdujo a la ciencia de datos como una disciplina independiente, extendiendo el campo de la estadística para incluir los avances en computación con datos en su artículo *Ciencia de datos: un plan de acción para expandir las áreas técnicas del campo de la estadística*. Cleveland estableció seis áreas técnicas que en su opinión conformarían al campo de la ciencia de datos: investigaciones multidisciplinarias, modelos y métodos para datos, computación con datos, pedagogía, evaluación de herramientas, y teoría.

En abril del 2002, el *Council for Science: Committee on Data for Science and Technology* (CODATA) empezó la publicación del *Data Science Journal*, enfocada en problemas como la descripción de sistemas de datos, su publicación en Internet, sus aplicaciones y problemas legales. Poco después, en enero del 2003, la Universidad de Columbia empezó a publicar *The Journal of Data Science*, la cual ofreció una plataforma para que todos los profesionales de datos presentaran sus perspectivas e intercambiaran ideas (Wikipedia, 2016).

2.2.2. El Científico de Datos y su Rol como Investigador

Las personas que se dedican a la ciencia de datos se les conoce como científico de datos. El proyecto *Master in Data Science* define al científico de datos como una mezcla de estadísticos, computólogos y pensadores creativos, con las siguientes habilidades:

- Recopilar, procesar y extraer valor de las diversas y extensas bases de datos.
- Imaginación para comprender, visualizar y comunicar sus conclusiones a los no científicos de datos.
- Capacidad para crear soluciones basadas en datos que aumentan los beneficios, reducen los costos.

Los científicos de datos trabajan en todas las industrias y hacen frente a los grandes proyectos de datos en todos los niveles. La definición mas famosa de las habilidades que componen a un científico de datos se han atribuido al Dr. Nathan Yau, quien precisó lo siguiente:

> el científico de datos es un estadístico que debería aprender interfaces de programación de aplicaciones (API), bases de datos y extracción de datos; es un diseñador que deberá aprender a programar; y es un computólogo que deberá saber analizar y encontrar datos con significado.

En la tesis doctoral de Benjamin Fry (Fry, J., 2004) explicó que el proceso para comprender mejor a los datos comenzaba con una serie de números y el objetivo de responder preguntas sobre los datos, en cada fase del proceso que él propone (adquirir, analizar, filtrar, extraer, representar, refinar e interactuar), se requiere de diferentes enfoques especializados que aporten a una mejor comprensión de los datos. Entre los enfoques que menciona Fry están: ingenieros en sistemas, matemáticos, estadísticos, diseñadores gráficos, especialistas en visualización de la información y especialistas en interacciones hombre-máquina, mejor conocidos por sus siglas en inglés "HCI" (Human-Computer Interaction). Además, Fry afirmó que contar con diferentes enfoques especializados lejos de resolver el problema de entendimiento de datos, se convierte en parte del problema, ya que cada especialización conduce de manera aislada el problema y el camino hacia la solución se puede perder algo en cada transición del proceso.

2.2.3. La Ciencia de Datos como Herramienta Predictiva

Uno de los enfoques principales de la ciencia de datos es el procesamiento de datos estructurados o no estructurados para la obtención de conocimiento. Es importante destacar que la ciencia de datos trabaja en condiciones especiales que la definen de otras disciplinas (como por ejemplo, la inteligencia de negocios).

- Trabaja en datos incompletos; es muy probable que hasta un setenta por ciento del tiempo de un científico de datos se utilice en recopilar y curar datos aparentemente no-relacionados, incompletos, o altamente dispersos.
- Los datos suelen estar desordenados; las fuentes de los datos a utilizar pueden ser de fuentes diversas y formatos diferentes, especialmente si estos datos provienen del Internet
- Analiza los datos para ver qué información obtiene; la exploración de datos no tiene garantía de hallazgo alguno como procedimiento científico, a diferencia de la inteligencia de negocios que opera sobre juegos de datos donde siempre hay certeza de al menos una conclusión
- Los hallazgos impulsan decisiones sobre operaciones y productos; no solo de negocios sino dentro del mundo de la investigacion de otras disciplinas, tales como la genética, biología, inteligencia artificial, etc.

Lo que distingue a la ciencia de datos de sus mismas técnicas y metodologías es su objetivo central de desplegar modelos efectivos para la toma de decisiones en un medio ambiente de producción. Así es una disciplina que que administra el proceso de transformar hipótesis y data en predicciones accionables (Zumel, N. y Mount, J., 2014). Los objetivos de predicción mas comunes incluyen la predicción de quien ganara una elección política presidencial, que productos se venderán mejor juntos, que créditos resultaran en moratoria, y cual pagina web el consumidor hará clic en la próxima hora.

2.2.4. Diseño de un Estudio de Ciencia de Datos

El científico de datos es responsable de guiar el proyecto de ciencia de datos de comienzo a fin. El exito de un proyecto de ciencia de datos no se da por la utilización de alguna herramienta en particular, sino de tener goles cuantificables, buena metodología, interacción interdisciplinaria, y un flujo de trabajo adecuado. Hay seis pasos principales en el diseño de un proyecto de ciencia de datos (Zumel, N. y Mount, J., 2014).

1. **Definir el objetivo:** El primer paso en un proyecto de ciencia de datos es definir un objetivo medible y cuantificable. En esta etapa se trata de aprender todo lo posible sobre el contexto del problema. Un objetivo concreto incluye condiciones firmes para definir el éxito de la solución y criterios de aplicación.

2. **Recopilar y administrar la data:** El segundo paso incluye identificar los datos necesarios para alcanzar los objetivos propuestos, explorar dicha data, y acondicionarla para hacerla aplicable al análisis. Esta etapa suele ser una de las más intensiva en el uso de tiempo y recursos y es también la más importante. El investigador debe contestar las preguntas más críticas. ¿Qué datos se tienen disponibles? ¿Cuáles de estos datos son los necesarios para resolver el problema? ¿La data disponible es suficiente o se necesita más información? ¿La calidad de la data es óptima?

3. **Construir el modelo de predicción:** La etapa de construcción del modelo es aquella donde se extrae información relevante de los datos para alcanzar el objetivo de estudio. Dado que muchas de las técnicas y procedimientos de modelos hace uso de suposiciones iniciales sobre la distribución de la data y sus relaciones, es muy probable que el investigador tenga que retroceder a la fase anterior, curar la data, y volver a la etapa de modelo en varias interacciones.

4. **Evaluar y criticar el modelo:** Una vez se obtiene el modelo, es necesario ver si se ajusta al problema en cuestión. ¿Es lo suficientemente preciso? ¿Su uso se generaliza bien? ¿Su desempeño es mejor que las herramientas disponibles existentes? Los resultados del modelo (coeficientes, agrupaciones, reglas, etc.) hacen sentido dentro del contexto del problema?

5. **Presentar los hallazgos y documentar:** A partir del momento que el investigador aprueba el modelo de datos, es importante la presentación de los mismos con el rigor científico esperado por aquellos que tienen implicación o serán evaluadores del proyecto de investigación. ¿

6. **Implementar el modelo en producción:** Muchos de los modelos de datos utilizados en la ciencia de datos deberán ser implementados como herramientas en la vida real. A esto se le conoce como despliegue en producción y tiene implicaciones de implementación que muchas veces salen de las manos del científico de datos y hacia el equipo de ingeniería.

2.2.5. Tareas Comunes en la Ciencia de Datos

Hemos hablado de la ciencia de datos y su carácter predictivo. Las tareas mas comunes para lo cual se utiliza la ciencia de datos son las siguientes.

- **Clasificación:** Decidir si algo pertenece a una categoría u otra
- **Puntuación:** Predecir o estimar un valor numérico, tal como lo es un precio o la probabilidad de un fenómeno
- **Ranking:** Aprender a ordenar objetos por preferencias
- **Agrupamientos:** Agrupar objetos en grupos de características homogéneas
- **Relaciones:** Encontrar relaciones o causas potenciales de efecto tal cual se ven en la data
- **Caracterizaciones:** Utilización general de visualizaciones y reportes de la data

2.3. Aprendizaje Automatizado

2.3.1. Introducción al Aprendizaje Automatizado

Es interesante que los métodos de aprendizaje automatizado proliferaron de forma paralela al concepto de ciencia de datos, y solo fueron absorbidos por esta en los últimos diez años. Alpaydim nos describe el aprendizaje automatizado como la programación de computadoras para optimizar un criterio de desempeño utilizando datos o experiencia pasada (Alpaydim, E., 2010). Tom Mitchell respeta este concepto al describir el aprendizaje automatizado como "... la construcción de programas computacionales que aprenden con la experiencia..." (Mitchell, T., 1997, pág. XV). Solo Peter Harrington utiliza una descripción mucho más simplista al determinar que "El aprendizaje automatizado es la extracción de información de la data." (Harrington, P. 2012, pág. 5).

Estudiar los procedimientos de aprendizaje automatizado equivale a estudiar tres temas principales que los componen.

- Diseño del estudio: conjuntos de entrenamiento y conjuntos de predicción
- Problemas conceptuales: error fuera de la muestra, curvas ROC

- Implementación práctica: en este caso en particular, un tema que se cubrirá con la biblioteca Caret

Todo el mundo predice todo tipo de aseveraciones, desde el resultado de una elección presidencial hasta el partido de fútbol del domingo de una liga en particular. Pero en el sentido estricto de la palabra, ¿qué significa predecir? En nuestro contexto científico, definiremos el acto de predecir como el resultado de utilizar la probabilidad y muestreo para la selección de un conjunto de entrenamiento, el cual utilizaremos para construir las características de diseño de una función de predicción. La función utilizará dichas características para generar nuevas predicciones. Los componentes para la selección adecuada de variables de predicción son los siguientes:

TO DO: Agregar el siguiente esquema Pregunta ¿Datos ¿Atributos ¿Algoritmo ¿Parámetros ¿Evaluación

Un ejemplo muy común utilizado generalmente para explicar el uso del aprendizaje automatizado es la detección de correo chatarra, también conocido como spam. Podemos utilizar atributos cuantitativos de los mensajes, por ejemplo la frecuencia de ciertas palabras, para que un modelo se entrene y pueda predecir dentro de ciertos rangos de certeza si un correo cualquiera es o no spam.

Importancia Relativa de Los Pasos

Hay una secuencia de pasos importante para la consecución de modelos de aprendizaje automatizado coherentes.

TO DO: Agregar el siguiente esquema como gráfico Pregunta : Data : Atributos : Algoritmos

La combinación de algunos datos y un deseo extremo de conseguir una respuesta no nos asegura que una razonable pueda extraerse de un cuerpo cualquiera de información (Tukey, 1977). También es útil recordar que la calidad de los datos que ingresan al conjunto de entrenamiento tienen un efecto sobre el resultado del modelo. Datos que no son útiles no aportan nada. Es mucho mejor que la data sea curada y organizada de manera que tenga alta relevancia al tema de estudio.

Los buenos atributos son aquellos que comparten las siguientes características:

1. ayudan a comprimir la data
2. retienen el mayor volumen de información relevante

3. son creados basados en un modelo experto del modelo a aplicarse

No es fácil hacer una buena selección de atributos que mas adelante se convertirán en variables de predicción. Los errores mas comunes son los siguientes.

1. tratar de automatizar la selección de atributos
2. no prestar la atención necesaria a las variaciones y particularidades de los datos
3. desechar información importante innecesariamente

En este sentido los algoritmos importan mucho menos que la selección y curación de la data a utilizar. Los mejores métodos de aprendizaje automatizado reúnen una serie de características que los hace justamente sobresalir del montón. Las características en mención son las siguientes:

1. Interpretable:
2. Simples:
3. Precisos:
4. Rápidos (de entrenar y evaluar):
5. Escalables:

La predicción de modelos se basa mucho en el arte de compensar beneficios versus necesidades.

1. interpretación de los datos vs. precisión
2. velocidad vs. precisión
3. simplicidad vs. precisión
4. modelos escalables vs. precisión

A pesar de tener que sopesar la mejor forma de compensar todas estas variables, la interpretación es muy importante y debe conservar su lugar, ya que poco sirve un modelo rápido y preciso que no se puede interpretar. Muchos autores otorgan un segundo lugar de importancia a lo escalable del modelo. Se han dado casos donde modelos muy precisos no se han podido

poner en producción por la complejidad de escalar el algoritmo. El caso mas mencionado es el premio NETFLIX, el cual otorgo un millón de dolares al equipo con el mejor modelo de predicción de gustos de sus clientes, solo para luego llegar a la conclusión que el mismo era demasiado complejo y lento de escalar en producción y archivarlo (Techdirt, 2012).

2.3.2. Métodos Supervisados y No-Supervisados

Para los autores Hastie, Tibshirani, y Friedman el aprendizaje supervisado intenta aprender una función f de predicción a través del uso de uso juegos de datos de entrenamiento en forma de muestras del total de los datos disponibles. El uso de datos de entrenamiento le permite al sistema aprender y minimizar el error del modelo de predicción [Hastie et al., 1997].

Harrington nos da una explicación más sencilla del término, al aclarar que el aprendizaje supervisado es aquel que le pide al computador aprender de los datos utilizando una variable específica como objetivo. Esto reduce la complejidad de algoritmos y patrones que se deben derivar de la muestra de datos [Harrington, 2012].

El profesor Alpaydin agrega que el aprendizaje supervisado tiene como objeto aprender un mapeo de los elementos de entrada a los de salida, teniendo en cuenta que los valores correctos de estos últimos están dados por el supervisor [Alpaydin, 2010].

2.3.3. Error Muestral y Error Fuera de Muestra

El siguiente concepto es fundamental dentro de la teoría de aprendizaje automatizado, y la terminología puede diferir un poco de los términos establecidos en la estadística inferencial.

- **Error dentro de la muestra:** es el margen de error que se obtiene al utilizar el juego de datos de entrenamiento en la construcción del modelo de predicción. También se conoce como error de re-substitución
- **Error fuera de muestra:** es el margen de error que se obtiene cuando se aplica el modelo de predicción a un nuevo juego de datos. También se lo conoce como error de generalización.

En este punto debemos aclarar cuales son las ideas principales en las que hay que enfocarse.

1. Principalmente estamos interesados mucho mas en el error de generalización - el que se obtiene al aplicar un nuevo juego de datos al modelo de predicción - que del margen de error de resubstitución.

2. El error de resubstitución siempre va a ser menor que el error de generalización

3. La razon por la cual se da este fenómeno (que el error de resubstitución sea menor que el error de generalización) es el efecto de sobreajuste. El algoritmo se está ajustando de más a los datos.

La data en la ciencia de datos tiene dos partes: señal y ruido. El objetivo del modelo de predicción es el de predecir la señal. Siempre se puede diseñar un modelo perfecto que capture tanto la señal como el ruido. Pero dicho modelo no se desempeñará bien en juegos de datos nuevos.

El efecto de sobreajuste se como la creación de un modelo optimista a partir del juego de datos de entrenamiento. Los métodos que utilizamos buscan interpretar los datos de tal manera que no solo se ajustan a la señal sino al ruido de los mismos. Por esa razón el margen de error de resubstitución (error dentro de la muestra) es tan bajo pero cuando se prueba el mismo modelo entrenado en un juego de datos externo el margen de error generalizado (fuera de la muestra) crece. Se ha comprobado que los errores por sobreajuste ocurren más en modelos complejos que en modelos sencillos. La razón es que muchas veces el modelo complejo es precisamente más complicado para ajustarse mejor a la señal de los datos, sin que estos ajustes sean necesarios - o precisos - al momento de cambiar del juego de datos.

2.3.4. Diseño de un Estudio de Aprendizaje Automatizado

El diseño de una investigación de ciencia de datos tiene seis pasos. El diseño del estudio de un problema de aprendizaje automatizado debe verse como el diseño de la fase de modelo (paso tres) mucho más detallado para no confundirlos. La metodología recomendada por el Dr. Jeff Leek (Leek, J. 2015) recomienda los siguientes seis:

1. Definir el margen de error deseado

2. Dividir la data en juegos específicos de entrenamiento, evaluación y validación (opcional)

3. En el juego de entrenamiento, seleccionar atributos y utilizar validación cruzada

4. En el juego de entrenamiento, seleccionar la función de predicción; utilizar nuevamente validación cruzada

5. si no se utilizo validación cruzada, aplicar prueba 1X al juego de evaluación

6. si se utilizo validación cruzada, aplicar prueba al juego de evaluación, refinar el algoritmo, y luego volver a someter 1X al juego de validación

A pesar de que no tiene una comprobación científica, la comunidad siempre aconseja evitar las muestras pequeñas de la misma forma que se evitan en la estadística clásica. Una pregunta válida es cuanto de los datos disponibles se deben destinar al juego de entrenamiento, cuantos al juego de validación y cuantos al juego de evaluación. Zumel y Mount (Zumel, N., Mount, J., 2014) consideran un modelo sencillo de división con 90 % de los datos destinados al entrenamiento de modelos y el 10 % restante a la evaluación. Sin embargo Leek (Leek, J. 2015) en su libro *Data Style* nos da un juego de reglas mas comprensivas de como distribuir los datos segun el volumen de los mismos.

A. Si el volumen de datos es grande

- 60 % para el juego de entrenamiento
- 20 % para el juego de evaluación
- 20 % para el juego de validación

B. Si el volumen de datos es mediano

- 60 % para el juego de entrenamiento
- 40 % para el juego de evaluación

C. Si el volumen de datos es pequeño

- entrenar sobre el 100 % de los datos
- utilizar validación cruzada sobre el mismo juego que se entrenó
- no ocultar el hecho hacer alusión en la investigación de la muestra poco representativa

La tentación de utilizar el juego de datos de validación y/o evaluación es muy grande para todos los científicos de datos noveles. Sin embargo la literatura concuerda en que no se debe utilizar la evaluación sino hacia el final del proceso.

La selección de que datos en particular deben elegirse en cada grupo debe ser aleatoria, con un porcentaje definido en cada uno pero total certeza de que no hay parcialidad en la selección. En el caso del lenguaje R, la biblioteca CARET tiene incorporada funciones para garantizar asignaciones de datos a los grupos de entrenamiento y evaluación totalmente aleatorios. Los juegos finales de datos deben reflejar sin embargo las mismas estructuras del problema. Un claro ejemplo son las series de tiempo (Hyndman, R., 2012) en las cuales los datos tiene un componente de tiempo que denota un orden en especial. De estos grupos debe seleccionarse muestras aleatorias pero representativas de los periodos de tiempo a fin de tener sentido. A su vez cada sub-muestra debe reflejar el mayor grado de diversidad posible. Esto se debe lograr con selección aleatoria pero a veces es difícil mantener dicho balance con la mezcla posible de atributos.

2.3.5. Tipo de Errores

El concepto de error en estadística es uno que embarca varias dimensiones. En lo que respecta al aprendizaje automatizado, no importa que tan grande sea la muestra ni que tan exacto sea el algoritmo, siempre cabe la probabilidad - aunque pequeña - que una predicción sea falsa a pesar de que arroja un resultado positivo. Podemos entonces dividir los tipos de errores según su predicción y verdadera naturaleza (Yakir, B. 2011).

En lineas generales diremos que un resultado es positivo si ha sido identificado como tal, y que es negativo si ha sido rechazado. De tal forma:

- **verdadero positivo:** es aquel que ha sido correctamente identificado
- **falso positivo:** es aquel que ha sido incorrectamente identificado
- **verdadero negativo:** es aquel que ha sido correctamente rechazado
- **falso negativo:** es aquel que ha sido incorrectamente rechazado

La combinación de los siguientes resultados nos permite medir estadísticamente variables pertinentes a los resultados del modelo. Estas variables se conocen como sensibilidad, especificidad, valor predictivo positivo, valor predictivo negativo, y exactitud.

Sensibilidad: La sensibilidad es la probabilidad que un fenómeno arroje un valor positivo cuando realmente lo es. Por ejemplo, un examen de una enfermedad da positivo cuando el paciente realmente esta enfermo de dicho padecer. Podemos expresar la formula como un cociente de la siguiente forma:

$$sensibilidad = \frac{VP}{(VP + FN)} \tag{2.1}$$

Especificidad: La especificidad es la probabilidad que un fenómeno arroje un valor negativo cuando realmente no se encuentra presente (o sea es una predicción negativa cuando la realidad también es negativo). Por ejemplo, un examen de embarazo que da negativo cuando la paciente no esta embarazada. Podemos expresar la formula como un cociente de la siguiente forma:

$$especificidad = \frac{VN}{(FP + VN)} \tag{2.2}$$

Valor Predictivo Positivo: El valor predictivo positivo es la probabilidad de que un fenomeno este presente cuando la predicción arroja un valor positivo. Por ejemplo, la probabilidad de que un paciente tenga diabetes cuando el examen arroja positivo. Podemos expresar la formula como un cociente de la siguiente forma:

$$valor\ predictivo\ positivo = \frac{VP}{(VP + FP)} \tag{2.3}$$

Valor Predictivo Negativo: Lo opuesto del valor predictivo positivo, es la probabilidad de que una prediccion arroje negativo cuando el fenómeno no este presente. Por ejemplo, la probabilidad de que un paciente no se le detecte diabetes cuando en la vida real no la tiene. Podemos expresar la formula como un cociente de la siguiente forma:

$$valor\ predictivo\ negativo = \frac{VN}{(VN + FN)} \tag{2.4}$$

Exactitud: Quizás el mas sencillo de percibir de forma natural, la exactitud es simplemente la probabilidad de una prediccion correcta. Podemos expresar la formula como un cociente de la siguiente forma:

$$exactitud = \frac{VP + VN}{(VP + FP + VN + FN)} \tag{2.5}$$

Midiendo Error en Data Continua

Para data continua, de naturaleza numérico, las dos formas de medir el error mas comunes en aprendizaje automatizado son el error cuadrático medio y la raíz error cuadrático medio.

La raíz error cuadrático media es utilizada con frecuencia para medir la diferencia entre valores (de una muestra y valores de una población) predicha por un modelo o un estimador y los datos observados en la realidad. Este valor representa la desviación estándar de la muestra entre los valores predecidos y los valores observados. Las diferencias individuales entre estas dos medidas se conocen como residuos si son extraídos de la muestra, y errores de predicción si son calculados fuera de muestra.

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(prediccion_i - observado_i)^2} \tag{2.6}$$

2.3.6. Sobreajuste

En aprendizaje automatizado, el sobreajuste (también es frecuente emplear el término en inglés overfitting) es el efecto de sobre-entrenar un algoritmo de aprendizaje con unos ciertos datos para los que se conoce el resultado deseado. Daroczi define el sobreajuste como la descripción del modelo en conjunto con el ruido aleatorio de la muestra en vez de solo el fenómeno generador de datos [Daroczi, 2015]. El sobreajuste ocurre, por ejemplo, cuando el modelo tiene más predictores de los que puede acomodar la muestra de datos.

Según Zumel y Mount, una de las señales de sobreajuste más sencillas de detectar se da cuando un modelo tiene un excelente desempeño en el juego de datos que se entrenó, pero uno muy malo en un juego de datos nuevo [Zumel and Mount, 2014]. Esto es causa y efecto de memorizar la data de entrenamiento en vez de aprender reglas generales de la generación del patrón.

2.3.7. R y la Biblioteca CARET

La biblioteca *CARET* (nombre extraído de Classification And Regression Training) es una libreria de funciones en R para optimizar el proceso de crear modelos predictivos. El paquete contiene herramientas para:

- segmentar juegos de datos
- preproceso de los datos
- seleccion de predictores
- optimizacion del modelo utilizando reconfiguracion de muestras

- estimacion de la importancia de la variable

El paquete esta mantenido en GitHub bajo la administración del Doctor en Estadística Max Kuhn.

2.4. Pronosticando Valores con Regresión

2.5. Introducción

Zumel y Mount definen *métodos funcionales* como aquellos modelos mejor adaptados para las tareas de clasificación y puntuación. Estos métodos son aquellos que pueden aprender de un modelo que es una función continua de sus entradas. Este tipo de métodos es especialmente útil cuando el investigador no solo quiere un valor de predicción, sino también medir la relación entre variable de entrada y resultados [Zumel and Mount, 2014]. En este caso se utiliza la función como guía del resultado esperado o predicción. Downey describe la regresión lineal como aquella que está basada en modelos de funciones lineales [Downey, 2014]. Para Mann y Lacke la regresión lineal es aquella que se da como una función lineal entre dos variables, y la cual se puede dibujar en el plano cartesiano como una recta [Mann and Lacke, 2010].

La teoría detrás de la regresión lineal es bastante homogénea a través de todos los autores. Zumel y Mount describen la regresión lineal como el más común de los métodos de aprendizaje automatizado [Zumel and Mount, 2014]. Para los autores hay una probabilidad muy grande que el método funcione bien con el problema, y si no, es muy fácil verificar cual otro método probar como segunda opción. Para Daroczi, el énfasis está en los modelos de regresión multivariable (una extensión de la regresión lineal simple de un solo predictor y resultado) que construyen el camino para la predicción de fenómenos complejos en la naturaleza y negocios (Daróczi, G., 2015). Por su parte, Harrington resume los beneficios de la regresión lineal (Harrington, P., 2012) por la facilidad de interpretar los resultados y lo frugal en el uso de ciclos de computación (aunque puede ser menos útil si el fenómeno no es perfectamente lineal).

2.6. Regresión Lineal

La regresión lineal modela el valor esperado de una cantidad numérica (llamada la variable dependiente, explicada o regresando) en términos

de entradas categóricas y numéricas (llamadas las variables independientes, explicativas o regresores) [Zumel and Mount, 2014]. El profesor Yakir de la Universidad de Jerusalem explica que el método más utilizado para describir la relación entre dos variables es la regresión. En el caso particular de la regresión lineal, buscamos la relación lineal entre dos variables numéricas. Este tipo de regresión calza la data a una linea. La linea resume el efecto de la variable exploratoria sobre la distribución de la respuesta [Yakir, 2011]

La regresión lineal describe la tendencia lineal en la relación entre respuesta y una variable exploratoria. Esta tendencia responde a la ecuación lineal en la forma de:

$$y = a + bx$$

donde y y x son variables y a y b son coeficientes en la ecuación. El coeficiente a se llama la intercepción o punto de corte con el eje de las ordenadas, y el coeficiente b la pendiente.

Partiendo de la formula de la ecuación lineal, podemos expander los términos a la solución de una regresión lineal. La mayor parte de la teoría de esta sección sigue el desarrollo de la fórmula:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

de tal forma que:

Y_i: variable dependiente, explicada o regresando

$X_1, X_2, \cdots, X_i$: variables explicativas, independientes o regresores

$\beta_0, \beta_1, \beta_2, \cdots, \beta_i$: parámetros, miden la influencia que las variables explicativas tienen sobre el regresor

donde:

β_0 es la intersección o término constante,

las $\beta_i (i > 0)$ son los parámetros respectivos a cada variable independiente,

y p es el número de parámetros independientes a tener en cuenta en la regresión [Yakir, 2011].

En términos generales, si suponemos que y_i es la cantidad numérica que uno desea predecir, y que x_i es un arreglo de datos (entradas) que se corresponden con los valores resultantes (salidas) de y_i, entonces la regresión lineal busca la función que calza tales valores de forma que:

$$y_i \to f(x_i) = \beta_1 x_{i,1} + \cdots + \beta_n x_{i,n} + \epsilon_i$$

Este último término ϵ_i representa lo que se llama el *error asintomático*, *perturbación aleatoria* o *ruido* que recoge todos aquellos factores de la realidad no controlables u observables y que por tanto se asocian con el azar, y es la que confiere al modelo su carácter estocástico. La sumatoria para todos los valores del error asintomático siempre promedia cero, y sus valores siempre están no correlacionados con los de x_i o y_i [Zumel and Mount, 2014].

Bajo estas premisas, la regresión lineal es implacable buscando los coeficientes. De existir alguna combinación ventajosa o cancelación de variables, la regresión lineal las encontrará. Lo único que no hace la regresión lineal es transformar las variables para que sean lineales (en muchos casos no lo son, por eso las técnicas de investigación exploratívas son importantes en la ciencia de datos previo a la búsqueda del modelo).

Los coeficientes de correlación miden el grado de relación entre variables y el signo de la misma, pero no la pendiente de la función. Hay muchas formas de medir la pendiente, pero la metodología más utilizada en el *Método de Mínimos Cuadrados* [Downey, 2014]. Un *ajuste lineal* es una linea cuya intención modelar la relación entre dos variables. Un *ajuste de mínimos cuadrados* es uno que minimiza el error cuadrático promedio entre la linea y los datos. Bajos las premisas de nuestra función lineal para cada valor y_i hay una ecuación correspondiente resultante de la suma del valor del término constante más el producto de la pendiente por el valor independiente. Pero a menos que la relación sea perfecta, siempre habrá una desviación entre el valor de predicción (o sea $\hat{y}_i$) y el valor real (y_i). A esta desviación se le denomina *residuo*.

El método de mínimos cuadrados busca minimizar la suma de los cuadrados de los residuos. La idea básica es que la función tiene el ajuste óptimo cuando la suma de los residuos es mínima (hay menor error en todos los puntos de la predicción). Es común para los estudiantes perder de vista porqué los valores tienen que estar elevados al cuadrado. Existen cuatro razones principales citadas por Dowen [Downey, 2014]:

1. elevar los términos al cuadrado tiene el efecto de evaluar de igual forma valores residuales positivos y negativos

2. elevar los términos al cuadrado tiene el efecto de darle mayor peso a los residuos de mayor valor, pero no tanto que desequilibren la fórmula

3. si los residuos no tienen correlación y están distribuidos de formal normal, con $\mu = 0$ y varianza desconocida pero constante, entonces el método de mínimos cuadrados y su ajuste de datos es también el mejor estimador para el término constante y la pendiente.

4. los valores del término constante y la pendiente que minimizan el cuadrado de los residuos pueden ser calculados con eficiencia dentro del ambito de la computación

Dowen agrega con el advenimiento de sistemas cada vez más poderosos, matemáticamente el método de mínimos cuadrados puede ser el óptimo, pero no necesariamente el más ventajoso [Downey, 2014].

2.7. Regresión Múltiple

Para medir el peso de la relación entre regresando y regresor, la regresión de un solo término es suficiente. Pero en la vida real, los problemas que el investigador quiera resolver son más complejos y requieren modelos más robustos. Los modelos de regresión en general conllevan la intención de medir la relación entre regresor y regresando controlando por otras variables, llamadas variables de confusión.

> Una *variable de confusión* es una tercera variable que sesga (aumenta o disminuye) la asociación entre dos variables de un modelo de regresión. Las variables de confusión siempre están asociadas con las variables dependientes e independientes [Daroczi, 2015]

Una de las ideas detrás de un modelo de regresión es mantener los valores de las variables de confusión fijos, para controlar su efecto sobre el resto del modelo. Variables de confusión que son candidatos en potencia se agregan al modelo como variables de regresión, y el coeficiente de regresión del modelo (el *coeficiente parcial*) mide su efecto ajustado a otras variables de confusión [Daroczi, 2015].

La formula de regresión múltiple no difiere mucho de la fórmula presentada para regresión lineal:

$$Y_i = \beta_0 + \beta_1 X_i + \cdots + \beta_n X_i + \epsilon_i \tag{2.7}$$

En cierta forma la regresión múltiple es la metodología más aplicable a la resolución de problemas de predicción en sistemas complejos que atribuyen su resultado a más de una variable de fuerza. La idea central de la variable de confusión es evitar resultados de correlación entre dos variables que pueden ser matemáticamente correctos pero no tener mucho sentido en la vida real. Por ejemplo, en un estudio de ataques cardíacos, el investigador puede tener la variable de peso controlada por la variable de altura, o el nivel de ejercicio del paciente, etc.

2.8. Presunción del Modelo

Los modelos de regresión con técnicas de estimación conocidas hacen un número de presunciones sobre la variable de resultado, las variables de predicción, y sobre su relación [Daroczi, 2015]. Estas presunciones se resumen en cinco puntos principales.

1. Y es una variable continua, o sea, no es binaria, nominal u ordinal.

2. Los residuos son estadísticamente independientes. Una violación de esta premisa ocurre cuando se utiliza análisis de tendencia. Dado que los años consecutivos no son independientes el uno del otro, los errores tampoco lo serán. Ende la necesidad de descomponer series de tiempo al momento de analizarlas.

3. Existe una relación lineal estocástica entre cada valor de Y y su correspondiente X. Una violación de esta premisa ocurre cuando la relación no es exactamente lineal, sino que es una desviación de una tendencia lineal.

4. Y tiene una distribución normal (ajustando cada valor de X fijo). Esto es necesario para poder hacer inferencias de la regresión.

5. Y tiene el mismo valor de varianza, más allá del valor fijo de las $X's$. Esto es necesario para poder hacer inferencias de la regresión, y se conoce como el concepto de *homocedasticidad*. Si se viola la presunción de homocedasticidad, tenemos el efecto de *heterocedasticidad*.

Dentro de la ciencia de datos, los investigadores están más interesados en los efectos de estas presunciones sobre el ensamblaje de un modelo que la teoría en si. Si alguno de estas presunciones falla, una solución probable es buscar valores extremos (outliers). Dentro del lenguaje R, es muy sencillo

verificar si un modelo de regresión lineal cumple con las cinco condiciones utilizando la función `gvlma` del paquete con el mismo nombre.

2.9. Solución a un Problema de Regresión Lineal con Aprendizaje Automatizado

La biblioteca CARET nos permite contar con funciones previamente diseñadas en R para la resolución de problemas de regresión lineal en R utilizando aprendizaje automatizado. Esto incluye extraer un modelo de regresión lineal de un juego de datos especificando que variables utilizar como dependiente e independientes.

2.10. Calce de los Datos

Un modelo de regresión lineal nos devuelve una linea de tendencia que es la mejor linea que se ajusta a los datos. Pero por ser la mejor linea no necesariamente tiene un ajuste que se considere bueno. El significado de los parámetros de regresión se obtiene a través de una prueba de hipótesis, la cual postula que el parámetro en cuestión tiene un valor de cero. R devuelve una prueba de Test F por cada parámetro, un valor de significancia que sirve para evaluar la regresión. Un valor p de menos de 0.05 puede ser interpretado como "la linea de regresión es significativa". De otra manera no el modelo no tiene mucho valor matemático [Daroczi, 2015].

Sin embargo, aún si el valor de la prueba F arroja un valor significativo, no se puede por esto decir mucho sobre el ajuste de la linea de regresión. El error del ajuste lo caracteriza mejor los residuos. Hay varias maneras de medir la calidad de un modelo lineal, o la *bondad de ajuste*. Uno de los más sencillos es medir la desviación estándar de los residuos. Para cualquier predicción utilizando regresión lineal, la desviación estándar de los residuos es equivalente al error promedio cuadrático de los mismos [Downey, 2014].

Dos medidas muy utilizadas son el Coeficiente de Correlación de Pearson y el Coeficiente de Determinación.

2.10.1. R2 - Coeficiente de Determinación

La primera pregunta sobre el modelo de regresión es su calidad. ¿Qué tan bien la variable independiente explica la variable dependiente en el modelo de regresión? El *coeficiente de determinación* es uno de los conceptos que explica mejor esta pregunta [Mann and Lacke, 2010].

En estadística, el coeficiente de determinación, denominado R^2 y pronunciado R cuadrado, es un estadístico usado en el contexto de un modelo estadístico cuyo principal propósito es predecir futuros resultados o probar una hipótesis. El coeficiente determina la calidad del modelo para replicar los resultados, y la proporción de variación de los resultados que puede explicarse por el modelo.

Para la regresión lineal la fórmula está determinada por:

$$R^2 = \frac{\sigma_{xy}^2}{\sigma_x^2 \sigma_y^2} \tag{2.8}$$

Estos valores son ciertos para:

$$0 \leq r^2 \leq 1$$

El valor del coeficiente de determinación aumenta cuando se incluyen nuevas variables en el modelo, incluso cuando éstas son poco significativas o tienen poca correlación con la variable dependiente. Esto es un problema en ciencia de datos, donde por cada variable de predicción que se agregue se puede asumir - erróneamente - que el modelo mejora al conseguir un coeficiente de determinación mayor.

En términos didácticos podemos pensar del coeficiente de determinación como un valor entre cero y uno que explica el nivel de validez de la respuesta del modelo hacia el valor de predicción del mismo, o lo que es similar, el porcentaje de la variación total que el modelo explica [Leek, J., 2016].

2.10.2. R - Coeficiente de Correlación de Pearson

En estadística, el coeficiente de correlación de Pearson es una medida de la relación lineal entre dos variables aleatorias cuantitativas. A diferencia de la covarianza, la correlación de Pearson es independiente de la escala de medida de las variables.

De manera menos formal, podemos definir el coeficiente de correlación de Pearson como un índice que puede utilizarse para medir el grado de relación de dos variables siempre y cuando ambas sean cuantitativas. Mann explica que es muy útil pensar en el coeficiente de correlación como la medida de proximidad entre los puntos de datos en un plano cartesiano y la distancia a la linea de regresión [Mann and Lacke, 2010]. En un mundo perfecto los puntos del juego de datos se alinea de forma perfecta con la linea de regresión. La fórmula para el coeficiente de correlación es la siguiente:

$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2}} \tag{2.9}$$

El rango de comportamiento del coeficiente de correlación siempre fluctúa dentro de un rango definido de valores:

$$-1 \leq r \leq 1$$

Cuando el valor de r es 1, hablamos de una correlación perfecta positiva, pero tal caso se da muy pocas veces en el mundo real. Como regla práctica, un índice de correlación cerca a uno habla de una correlación fuerte, mientras que uno cercano a cero es señal de una correlación muy débil. El signo en el índice nos da la indicación de si la pendiente es positiva o negativa.

El cuadrado del indice de correlación es el indice de determinación de una regresión.

2.10.3. El valor de p

2.10.4. Seleccionando Variables de Predicción

2.11. Series de Tiempo

2.11.1. Introducción a las Series de Tiempo

Muchos autores han escrito sobre las series de tiempo, pero es difícil agregar al tema o discutir las ideas del profesor Robert Hyndman, uno de los expertos más respetados en la comunidad de la estadística por su trabajo en las series de tiempo. Hyndman extiende la teoría a las series de tiempo como elementos de pronostico y su relación con la regresión lineal (Hyndman, R., 2014). Desde el punto de vista técnico, Hyndman es el creador de varias bibliotecas de funciones de pronostico utilizando series de tiempo y ARIMA en lenguaje R. Dentro de la bibliografía, Daroczi es quien agrega detalles sobre la detección temprana de valores atípicos que pueden dificultar – y mucho – el análisis (Daróczi, G., 2015).

2.11.2. Pronóstico con Series de Tiempo

Los pronósticos con series de tiempo utilizan solamente la información disponible de la variable que se propone pronosticar, sin hacer intento alguno por descubrir los factores adicionales que condicionan su comportamiento. Por lo tanto se extrapolan las tendencias y patrones temporales, pero se

ignora toda la informacion adicional como pueden ser iniciativas de publicidad, actividad de la competencia, cambios en las condiciones económicas y otros [Hyndman and Athanasopoulos, 2014].

2.11.3. Patrones

Las series de tiempo pueden descomponerse según su patrón o tendencia en tres elementos que las componen [Velazco, M., 2017]. A saber:

1. Tendencia Secular: la tendencia secular o tendencia a largo plazo de una serie de tiempo es por lo común el resultado de factores a largo plazo.

2. Variación Estacional: Es el componente de la serie de tiempo que representa la variabilidad de los datos debido a la influencia de las estaciones.

3. Variación Irregular: Esta variación se debe a factores a corto plazo, imprevisibles, y no recurrentes que afectan la serie de tiempo.

2.11.4. Auto Correlación

De igual manera que una correlación mide la extensión de una relación linear entre dos variables, la autocorrelación mide la relación linear entre dos valores retrasados de series de tiempo [Hyndman and Athanasopoulos, 2014].

El valor de una autocorrelación para un r_k dado es:

$$r_k = \frac{\sum_{t=k+1}^{T}(y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^{T}(y_t - \bar{y})^2}$$

donde T es el valor de período temporal de la serie de tiempo.

El autor Daroczi agrega como metodología para la verificación de autocorrelación en un juego de datos (no solo una serie de tiempos, sino cualquier juego de datos espacial) el *Indice I de Moran* [Daroczi, 2015]. Dicho índice esta dado por la formula:

$$I = \frac{N}{W}\frac{\sum i \sum j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum i(x_i - \bar{x})^2}$$

2.11.5. Precisión del Pronostico

Existen dos formas que se utilizan comúnmente para medir la precisión del pronóstico de series de tiempo. Ambas están basadas en el error absoluto o error cuadrático [Hyndman and Athanasopoulos, 2014].

$$ErrorPromedioAbsoluto(MAE) = promedio(|e_i|)$$

$$ErrorPromedioCuadratico(RMSE) = \sqrt{promedio(e_i^2)}$$

La tendencia al comparar precisión en un solo juego de datos es utilizar el MAE ya que es mas sencillo y simple de entender.

2.11.6. Entrenamiento y Evaluación

Al igual que la mayoria de los metodos de aprendizaje automatizado, las series de tiempo se suelen entrenar y evaluar con juegos alternos de muestra de datos. El tamaño de cada uno varía con el investigador, pero en serie de datos el sistema operativo tiende a ochenta por ciento de la muestra para entrenamiento y veinte por ciento para evaluación [Hyndman and Athanasopoulos, 2014].

2.11.7. Descomposición de Series de Tiempo

La descomposición de las series de tiempo facilita el análisis y la investigación exploratoria de los datos. Una de las formas mas sencillas de lograr esto es la aplicación de promedios móviles, lo cual se facilita mucho en R con el uso de la función `decompose()` [Daroczi, 2015].

La descomposición por promedios móviles toma la forma siguiente:

$$s_t = \frac{1}{k} \sum_{n=0}^{k-1} x_{t-n}$$

Muchas series de tiempo no son aditivas sino multiplicativas, y con el paso del tiempo incrementan la amplitud de las fluctuaciones. Para tales series, la biblioteca de R tiene la función `stl()` que aplica descomposicion a base del método Loess. La aplicacion del método Loess y la transformación de la función a una logarítmica tiene el efecto de replicarla como aditiva [Viswanathan and Viswanathan, 2015].

2.11.8. Suavizar Exponencialmente con Holt-Winters

Es posible eliminar todos los efectos de estacionalidad en una serie de tiempo con la aplicacion del filtro Holt-Winters. Esto no solo resulta en una lectura mas clara de la tendencia secular de la serie, útil para pronósticos, sino que adicionalmente tiene el efecto de eliminar valores atípicos (outliers) en la misma [Daroczi, 2015].

2.11.9. ARIMA

Uno de los métodos mas populares para la descomposición de series de tiempo con periodos de doce meses es ARIMA, el cual fue desarrollado por el Buró de Censo de Estados Unidos [Hyndman and Athanasopoulos, 2014]. El método esta basado en la descomposición tradicional, pero tiene la ventaja de mantener la tendencia secular en todos los puntos de datos, y de permitir que la tendencia estacional varíe de a poco con el tiempo. Es también un método muy robusto.

2.11.10. Dickey-Fuller

Un componente importante de las series de datos es la detección de si son o no auto-regresivas (lo que determina mucho de su poder predictivo). La fórmula para la detección de series auto-regresivas es el test Dickey-Fuller, y la mejor bibliografía es el artículo científico escrito por ambos profesores en la revista especializada Econometrica (Dickey, D., y Fuller, W., 1981). A pesar de ser un artículo contemporáneo, la teoría detrás de la prueba Dickey-Fuller nos permite descartar series de tiempo no-regresivas con poco poder de predicción.

2.12. Modelos Ensamblados

El tema de modelos ensamblados es uno que por lo general se reserva más como técnica de composición que cómo teoría del aprendizaje automatizado.

2.12.1. Introducción

El uso de modelos ensamblados es en cierta forma la prueba final de la hipótesis de trabajo: la utilización de dos modelos entrecruzados cuyos resultados conforman una tabla temporal de valores esperados de los cuales se genera un nuevo modelo sintético de predicción más general y con mayor

capacidad de predicción en juegos de datos de validación cruzada. Este concepto es novel; Witten y Frank lo describen como combinación de métodos múltiples, y escriben: "... un enfoque obvio para hacer mejores decisiones es tomar el resultado de diferentes métodos y combinarlos..." (Witten, I. y Frank, E., 2005). Zhou nos describe que "... los modelos ensamblados que entrenan múltiples variables y luego las combinan para uso de entrenamiento, con el Boosting y el Bagging como representantes principales, representan lo más novedoso en el estado del arte de la ciencia de datos..." (Zhou, Z., 2012, pg. VII). De una manera un tanto más coloquial, Zhang y Ma describen el uso de modelos ensamblados con una analogía de la vida real, en la cual los pacientes buscan una segunda y hasta tercera opinión de expertos antes de someterse a una operación complicada (Zhang, C. Y Ma, Y., 2012). Curiosamente tanto Zhang, Ma y Zhou hablan de la combinación de métodos de regresión general con clasificadores, y solo Witten y Frank hablan de otras combinaciones (por supuesto, Witten y Frank comenzaban a escribir en los albores del ensamblaje de métodos, cuando los clasificadores no estaban tan de moda porque el análisis era mayoritariamente de números, algo que cambió con el avance de las redes sociales).

2.12.2. Combinando Métodos

La combinación de métodos es el ultimo paso en la estrategia de construcción de un sistema ensamblado de aprendizaje automatizado. La pregunta de que métodos combinar esta estrechamente relacionado con el tipo de juegos de datos y la solución que se busca alcanzar. Por ejemplo, alguno métodos de clasificación como los vectores de soporte solo devuelven valores discretos [Zhang and Ma, 2012]. De tal manera el uso de dos métodos alternos en uno ensamblado estará determinado por la forma final en que se ensamblan y el algoritmo final utilizado para la decisión de predicción. Tanto Polikar [Zhang and Ma, 2012] como Zhou [Zhou, 2012] citan como preferibles las metodologías de voto por mayoría, promedio, promedio ponderado, y ensamblaje infinito.

2.12.3. Diversidad

La diversidad de ensamblaje, o la diferencia entre diferentes métodos de aprendizaje, es un tema fundamental en el ensamblaje de métodos [Zhou, 2012]. Intuitivamente es fácil entender que para obtener una ventaja de la combinación, es necesario que los aprendizajes sean diferentes, de otra manera la ganancia en desempeño no seria marginalmente superior a los métodos por

separado [Zhou, 2012].

2.12.4. Bagging

La idea del *bagging* esta estrechamente ligada al *bootstrapping*, y determinada por la selección de múltiples muestras de datos generadas a través de *bootstrapping*, utilizadas para alimentar clasificadores, sobre cuyos resultados el método ensamblado puede votar [Daume, 2013].]

2.12.5. Boosting

El *boosting* es la técnica por la cual se toma un algoritmo de aprendizaje con malos resultados (técnicamente conocido como un clasificador débil) y se lo transforma en un clasificador fuerte. La forma en la cual funciona el *boosting* es que basado en un juego de datos y resultados pasados, va generando nuevas predicciones. Las predicciones con resultados aceptables se les pone menor peso y recursos, mientras que el algoritmo vuelve a iterar en aquellas predicciones con valores lejanos hasta que cobran fuerza [Daume, 2013]. Esta técnica recibe el nombre de **AdaBoost**, del ingles *adaptive boosting algotithm*. Esta fue una de las primeras técnicas practicas en la ciencia de datos.

Bibliografía

[Alpaydin, 2010] Alpaydin, E. (2010). *Introduction to Machine Learning 2nd Edition*. The MIT Press, Massachusetts, USA.

[Cárdenas, 2016] Cárdenas, A. O. (2016). *Economía Colombiana 5ta Edición*. ECOE Ediciones, Bogotá, Colombia.

[Carriello, 2010] Carriello, B. B. (2010). *Crisis Cambiarias en Países Emergentes*. Ediciones Uninorte, Barranquilla, Colombia.

[Daroczi, 2015] Daroczi, G. (2015). *Mastering Data Science with R*. Packt Publishing, Birmingham, UK.

[Daume, 2013] Daume, H. (2013). *A Course in Machine Learning*. University of Maryland, Maryland, USA.

[Downey, 2014] Downey, A. B. (2014). *Think Stats*. Green Tea Press, Massachusetts, USA.

[Harrington, 2012] Harrington, P. (2012). *Machine Learning in Action*. Manning Publications, Shelter Island, USA.

[Hastie et al., 1997] Hastie, T., Tibshirani, R., and Friedman, J. (1997). *The Elements of Statistical Learning*. Springer, Stanford, USA.

[Hyndman and Athanasopoulos, 2014] Hyndman, R. and Athanasopoulos, G. (2014). *Forecasting Principles and Practice*. Otexts, Melbourne Australia.

[Mann and Lacke, 2010] Mann, P. S. and Lacke, C. J. (2010). *Introductory Statistics*. Wiley, Willimantic, USA.

[Viswanathan and Viswanathan, 2015] Viswanathan, V. and Viswanathan, S. (2015). *R Data Analysis Cookbook*. Packt Publishing, Birmingham, UK.

[Yakir, 2011] Yakir, B. (2011). *Introduction to Statistical Thinking (With R, Without Calculus)*. The Hebrew University, Jerusalem, Israel.

[Zhang and Ma, 2012] Zhang, C. and Ma, Y. (2012). *Ensemble Machine Learning*. Springer, New York, USA.

[Zhou, 2012] Zhou, Z. (2012). *Ensemble Methods*. Chapman and Hall - CRC, Boca Raton, USA.

[Zumel and Mount, 2014] Zumel, N. and Mount, J. (2014). *Practical Data Science with R*. Manning, Shelter Island, USA.