UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

PREDICCIÓN DE MÚLTIPLES SERIES DE TIEMPO UNIVARIADAS A TRAVÉS DE DIVERSOS MODELOS PREDICTIVOS Y META-LEARNING APLICADO EN LA INDUSTRIA DEL RETAIL

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

LUIS ALBERTO GUTIÉRREZ GONZÁLEZ

PROFESOR GUÍA
MARCEL GOIC FIGUEROA

PROFESOR CO-GUÍA
ALEJANDRA PUENTE CHANDÍA

COMISIÓN
CHARLES THRAVES CORTÉS-MONROY

SANTIAGO DE CHILE

2020

RESUMEN DE LA MEMORIA PARA OPTAR AL TÍTULO DE: Ingeniero Civil Industrial
POR: Luis Alberto Gutiérrez González
FECHA: 14/07/2020
PROFESOR GUÍA: Marcel Goic Figueroa

PREDICCIÓN DE MÚLTIPLES SERIES DE TIEMPO UNIVARIADAS A TRAVÉS DE DIVERSOS MODELOS PREDICTIVOS Y META-LEARNING APLICADO EN LA INDUSTRIA DEL RETAIL

En la siguiente memoria, se aborda un problema de predicción de demanda de productos en la industria supermercadista usando avances metodológicos recientes. Esto es debido a que los modelos y metodologías utilizadas más recientes, son de a mediados del 2018. Además de esto, serán analizadas series de tiempo extraídas desde un retail nacional, estas series son interesantes y particulares de analizar debido a que el comportamiento de estas es bastante sucio, ya que se ven perturbadas por eventos no predecibles, ni tampoco descritos o medidos en el pasado, como lo son promociones únicas de fin de semana, o de horarios en particular. Por otro lado, también son afectadas por fechas festivas, feriados, situación país, aumento o disminución del valor de la moneda, etc. También se ven afectadas por algún desliz en la gestión de operaciones o logística de la empresa, esto se puede ver reflejado en no tener el stock del producto demandado, es decir, un quiebre de stock, generando un vacío en la historia del producto.

Dado estas condiciones particulares, se compararán un total de 7 modelos predictivos, sumado un clasificador capaz de elegir qué modelo utilizar para cada serie de tiempo. Es decir, no es necesario ejecutar los siete modelos, si no que este clasificador elige de antemano cuál es el mejor al entregarle la serie de tiempo con suficiente historia, reduciendo el tiempo de ejecución total en n veces en promedio, siendo n=7 la cantidad de modelos candidatos, debido a que al saber cuál es el mejor, es posible ahorrarse el tiempo de ejecución de los otros n-1 modelos. Se trabajará con un poco menos de tres años de historia, y con 5.000 series de tiempo de variado comportamiento entre ellas.

El objetivo es generar una predicción más precisa que un modelo de media móvil a dos semanas, lo que se hace en la actualidad en el centro de distribución del retail, el cual está encargado de reponer cada uno de los artículos a los distintos locales distribuidos en el país. Esto es para disminuir los costos asociados a dos grandes problemas. El primero, la subestimación de la demanda, lo que genera quiebres de stock en los distintos locales, perdiendo la oportunidad de haber vendido más unidades, y el segundo, es el costo asociado a una sobre estimación de la demanda de un producto, lo que lleva a un sobre stock en la bodega del local, implicando costos de almacenaje y espacio, pudiendo quitar espacio para productos que si están vendiéndose y en el caso final que no tengan rotación y estén estancados en la bodega, son liquidados a un precio incluso menor que al costo total generado, para liberar espacio y abrir cupo a nuevos artículos de temporada con una contribución mayor.

AGRADECIMIENTOS

A mis padres, que me apoyaron en todo momento, gracias por estar ahí siempre que los necesité, hoy soy quién soy gracias al esfuerzo y perseverancia que ustedes me han enseñado.

A mi hijo Gabriel Gutiérrez, que me ha mostrado el mundo con otros ojos, me llenas de alegrías con tu sonrisa y amor incondicional.

A Pablo Galaz, mi amigo que me acompañó en un sin fin de conversaciones, gracias por ser un apoyo constante, y alguien que siempre estuvo ahí a lo largo de este gran camino, sé que eres alguien en quién siempre podré confiar.

A Fernando Brierley, un amigo que no solo me ayudó a crecer profesionalmente, si no que me ayudo a crecer como persona, gracias por tus sabios consejos, siempre te he considerado un maestro para mí, y gracias por siempre ser una persona que se mueve desde el amor.

A todos esos amigos que fueron parte de mi día a día universitario e hicieron más alegre este viaje, gracias Vicho, Tomy, Aldo, Pablito, Chino, Mayer, Moscoso, Sole, Daco, Estaban, Miguelo, Juano, Dieguito, Edrian, Sofi, Anto. Muchas gracias por ser grandes personas que espero seguir viendo en las nuevas etapas que se nos vienen.

A David Nova, que me enseñó a ser profesional y persona al mismo tiempo, eres un gran amigo y espero seguir aprendiendo de ti.

Por último, gracias a todas las personas que me permitieron crecer en este largo camino, a docentes como Marcel Goic, Sebastián Balmaceda, Charles Thraves, Pedro Pineda. A las oportunidades que tuve de ser parte de equipos docente, donde muchas veces recibí más de lo que di, y a todos aquellos que regalaban una sonrisa día a día.

1 TABLA DE CONTENIDO

1 Tabla de contenido iii

1.1 Índice de tablas iv

1.2 Índice de ilustraciones v

2 Introducción 5

2.1 Características y antecedentes de la empresa 5

2.2 Descripción del problema 5

2.3 Alcances del trabajo 6

2.4 Objetivos 7

3 Desarrollo de la investigación 8

3.1 Exploración y visualización de los datos 8

3.1.1 Estructura y contexto de los datos 8

3.1.2 Visualización de la data 9

3.2 Metodología 11

3.3 Marco conceptual 12

3.3.1 Modelos predictivos 12

3.3.2 Modelo de clasificación 18

3.3.3 Métricas 19

3.4 Implementación y resultados de modelos predictivos 20

3.4.1 Análisis por criterio MAE 21

3.4.2 Análisis por mejor predicción 24

3.5 Conclusiones modelos predictivos 26

4 Predicción a través de meta-learning 28

4.1 Metodología 28

4.2 Generación de atributos para series de tiempo 28

4.3 Entrenamiento modelo random forest 31

4.3.1 CASO 1: Los tres modelos con mayor MAE 32

4.3.2 CASO 2: Dos mejores modelos según MAE 33

4.3.3 CASO 3: Todos los modelos según MAE 33

4.3.4 CASO 4: Todos los modelos entrenando solo con los dos mejores 35

4.4 Análisis de los resultados 36

5 Conclusiones 37

5.1 Discusión 38

6 Bibliografía 39

1.1 ÍNDICE DE TABLAS

Tabla 3-1: Evaluación Forecast a una semana 22
Tabla 3-2: Evaluación Forecast a dos semanas. 22
Tabla 3-3: Evaluación Forecast a tres semanas.....23
Tabla 3-4: Evaluación Forecast a cuatro semanas 23
Tabla 3-5: MAE modelos según horizonte de predicción 24
Tabla 3-6: Win rate de modelos según MAE para cada horizonte de pronóstico 26
Tabla 4-1: Matriz de confusión clasificador a una semana con los modelos de menor rendimiento 32
Tabla 4-2: Comparación entre modelo de clasificación a una semana con modelos de menor rendimiento 32
Tabla 4-3: Matriz de confusión clasificador a una semana con los modelos de rendimiento más alto 33
Tabla 4-4: Comparación entre modelo de clasificación a una semana con modelos de mayor rendimiento en set de testeo.....33
Tabla 4-5: Matriz de confusión clasificador a una semana con todos los modelos.....34
Tabla 4-6: Comparación entre modelo de clasificación a una semana con todos los modelos 34
Tabla 4-7:Comparación entre modelo de clasificación a una semana con modelos, entrenado con los mejores dos.....35

1.2 ÍNDICE DE ILUSTRACIONES

Ilustración 3-1: Promedio de ventas agregadas por sección 9
Ilustración 3-2: Ejemplos de comportamientos de productos.....10
Ilustración 3-3: Modelo Media móvil.....13
Ilustración 3-4: Modelo de Holt 13
Ilustración 3-5: Modelo ARIMA.....14
Ilustración 3-6: Modelo STLF.....15
Ilustración 3-7: Modelo TBATS.....16
Ilustración 3-8: Modelo NNetAR 17
Ilustración 3-9: Modelo ensamblado 18
Ilustración 3-10: Pronóstico de todos los modelos utilizados.....18
Ilustración 3-11: MAE por modelo a distintos horizontes. 21
Ilustración 3-12: Win rate de cada modelo según MAE para cada horizonte de pronóstico 25
Ilustración 4-1: Serie de tiempo Seasonal Strenght = 0.931 29
Ilustración 4-2: Serie de tiempo Seasonal Strenght = 0.254 29
Ilustración 4-3: Relación entre las características de las 5000 series de tiempo.....30
Ilustración 4-4: Serie con trend = 0.767, cálculo alterado por ventas 0 31

2 INTRODUCCIÓN

2.1 CARACTERÍSTICAS Y ANTECEDENTES DE LA EMPRESA

Cencosud es uno de los más grandes conglomerados de retail en América latina, con presencia en Argentina, Brasil, Perú y Colombia, donde se hacen presente con distintas unidades de negocio. Cuentan con un equipo de 140.000 colaboradores a través de los distintos países. Su Misión es trabajar día a día para llegar a ser el retail más rentable y prestigioso de América Latina, en base a su excelencia en calidad de servicio.

En base a la misión declarada, Cencosud muestra interés de mantenerse dentro de los líderes del retail, para esto, se considera vital el progreso de los proyectos relacionados con innovación y desarrollo. Una de las líneas presentes en I+D, está relacionada con la capacidad de analizar grandes volúmenes de datos, y extrayendo información relevante desde ellos. Además, es posible hacer predicciones de los futuros escenarios del retail aplicando herramientas de DataScience, vinculando las nuevas tecnologías con el negocio para poder competir fuertemente con el resto de los actores relevantes en el mercado que están desarrollando de igual manera esta área. Debido a esto, Cencosud crea la gerencia de Advanced Analytics en el 2016, encargada de liderar proyectos que utilicen estas nuevas tecnologías para mejorar los procesos de sus distintas unidades de negocio. Desde proyectos de pricing, hasta proyectos logísticos. Este tema de memoria se ubica en el proyecto "Allocation", el cuál nace por la necesidad de mejorar la calidad de la precisión de la reposición de los distintos productos en sus supermercados efectuada por el centro de distribución, con el objetivo de reducir los quiebres de stock de sus productos producidos por una subestimación de las ventas de ese periodo, y reducir el sobre stock, es decir, que las ventas sean considerablemente menores al stock del local, mientras que en otros existe falencia de este artículo, lo que obliga a liquidar el producto a un valor incluso menor a su costo.

2.2 DESCRIPCIÓN DEL PROBLEMA

Una de las decisiones que debe tomar un centro de distribución, es cuánto enviará de cada producto, a cada local, una decisión que debe considerar el nivel de ventas del local para cada producto para las próximas semanas, el nivel de stock que tienen para cada producto, el stock restante en el centro de distribución, el tamaño de la bodega de cada local, el tiempo de envío que varía dependiendo la distancia del local, la cantidad de veces que se puede mandar a cada local por semana, el stock de seguridad que se debe mantener para cada artículo, el volumen utilizado por caja de cada producto, entre otras variables consideradas en el proceso.

Esta decisión de cuánto reponer tiene más de una consecuencia, ya que en caso de mandar una cantidad menor a la que se podría vender durante las próximas semanas, se generará un quiebre de stock de ese artículo en el local, lo que primero disminuye la cantidad de ventas y ganancias netas,

y segundo genera una mala experiencia para los clientes, creando percepción de bajo nivel de servicio por parte del local. Por otro lado, reponer una cantidad por sobre la necesaria también es costoso, ya que mientras mayor sea el exceso, se generará un mayor sobre stock en la tienda, el cual puede ocupar un espacio importante en la bodega dependiendo del producto, lo que se acentúa si el producto es de baja rotación, impidiendo la entrada de nuevos productos a la bodega que si se podrían vender, es posible que la solución a este punto sea liquidar el producto a un precio menor que su costo, asumiendo el error cometido y las pérdidas asociadas. Otro punto importante es tomar una mala decisión sobre a qué local debo enviar el producto cuando quedan pocas unidades en el centro de distribución, es decir, mientras uno tiene sobre stock, existe otro local que vende ese producto en mayor nivel, generando un quiebre de stock. Este escenario es doloroso para los locales ya que no es posible reubicar productos que ya fueron enviados, debido a que este proceso tiene un alto nivel de costos.

En la actualidad, una vez que llegan los distintos artículos desde el extranjero al centro de distribución, existen personas dedicadas a monitorear el stock de los productos en los distintos locales, además de comunicarse con cada local para coordinar la cantidad a enviar en base a las condiciones del local y el centro de distribución. Luego de este intercambio de información, esta persona decide cuanto va a reponer, lo que depende de múltiples factores mencionados anteriormente. Este proceso en la realidad tiene bastantes inconvenientes, por ejemplo, no siempre es posible comunicarse con todos los locales, tomando decisiones con información incompleta.

Después de todo el proceso de recopilación de información, esta persona genera una predicción de las ventas para las próximas cuatro semanas con el fin de planificar cuánto debe mandar a cada local. Esta predicción se hace en base a un promedio de las ventas de las últimas dos semanas de cada artículo en cada local. Finalmente, la cantidad a reponer es calculada en base a la predicción calculada, al stock de los locales, el stock en tránsito (producto que está solicitado, pero aún no llega), y el stock del centro de distribución.

2.3 ALCANCES DEL TRABAJO

Dada la importancia de la precisión al momento de decidir cuánto se debe reponer de cada artículo en cada local, este trabajo se centra en estimar las ventas para distintos grupos de artículos, evaluando la metodología actual, y proponiendo nuevos métodos de pronóstico de mayor precisión, para así poder realizar una reposición más precisa. En este caso, se utilizarán distintos modelos matemáticos de predicción tanto contemporáneos como conservadores, además de metodologías novedosas como la predicción de series de tiempo univariadas a través de meta-learning. Esto será en un horizonte de cuatro semanas, situadas en septiembre del año 2019.

2.4 OBJETIVOS

El objetivo general de esta memoria de título consta:

"Desarrollar un modelo predictivo que permita generar un pronóstico más certero que la media móvil para productos de vestuario y juguetería de supermercados, para disminuir los costos logísticos de las áreas de operaciones causados por errores de pronóstico"

Objetivos específicos:

1. Generar un pronóstico más preciso que la media móvil en base a métricas definidas para asegurar un mejor resultado en el cálculo de la reposición, para el pronóstico a una, dos, tres, y cuatro semanas.

2. Realizar tabla comparativa de los modelos predictivos para entender la magnitud de la precisión de cada una de las alternativas/modelos propuestos.

3. Validar la calidad de la predicción de un modelo creado con técnicas de meta-learning, para explorar la posibilidad de un modelo diferente a los tradicionales que permita hacer más eficiente en tiempo y precisión el proceso de selección de modelos predictivos para series de tiempo.

3 DESARROLLO DE LA INVESTIGACIÓN

3.1 EXPLORACIÓN Y VISUALIZACIÓN DE LOS DATOS

Para encontrar una respuesta a cuál es la mejor solución para implementar, en primer lugar, los datos son explorados en profundidad para entender sus comportamientos, identificar los componentes principales de las diferentes series, y la existencia de irregularidades en los patrones observados o falta de data.

3.1.1 Estructura y contexto de los datos

Se tomaron para esta investigación un universo de 5000 series de tiempo univariadas, es decir, solo se cuenta con las ventas y la fecha en que estas fueron ocurridas. Estas ventas están agregadas a nivel semanal, esto es debido a que en general los productos tienen niveles de venta muy bajos a nivel diario. Las ventas comienzan desde enero 2017 (2017-01-02), y llegan hasta la cuarta semana de septiembre del 2019 (2019-09-23).

Los productos considerados en este estudio pertenecen a la categoría de VESTUARIO en su mayoría (90%), mientras que los restantes pertenecen a la categoría de JUGUETERÍA, esto es diseñado así para tener series de tiempo con distintos comportamientos, donde juguetería es estable a excepción de fechas como Navidad y Día del niño, y vestuario tiene un comportamiento de temporada/estacional. Además de esto, serán considerados como productos una agrupación de SKU's conveniente predefinido por el árbol de productos creado por la empresa, de forma tal que por ejemplo, el producto POLERAS M/C, son poleras manga corta de cierta marca sin importar la talla o color, se define de esta manera en la investigación por dos principales razones. La primera, está relacionada con la historia de los productos, a cada año salen nuevos artículos los cuales no tienen historia, lo que hace muy difícil entrenar buenos modelos y por ende generar una predicción precisa, mientras que cuando se agrupan, se cuenta con la historia de sus productos "hermanos", lo que permite entregar un pronóstico a pesar de que el artículo como tal no tenga historia, lo que se relaciona con la segunda razón, la cual es capturar la estacionalidad y tendencia de estos productos en su conjunto a través de los años, esto es solo asumiendo que tendrán un comportamiento similar entre ellos, es decir, la polera negra manga corta talla M y L tendrán una tendencia y estacionalidad similar, por lo que agruparlas hace sentido para ganar información, sobre todo en los casos con poca historia (menos de un año).

Es importante considerar que un mismo producto puede tener un comportamiento diferente entre un local y otro, considerando que en general ningún local es igual a otro, debido a que las personas que asisten a un local u otro son diversas entre ellas. Por esta razón definiremos que un producto

en dos locales será considerado como dos productos diferentes (uno puede vender mucho y otro no), esto debido a la diferencia entre el comportamiento de compra en los clientes entre locales.

3.1.2 Visualización de la data

A continuación, se visualiza el promedio de las ventas agregada de las 5.000 series de tiempo a analizar desde la semana 01 del 2017 hasta la semana 23 del 2019 separada por sección.

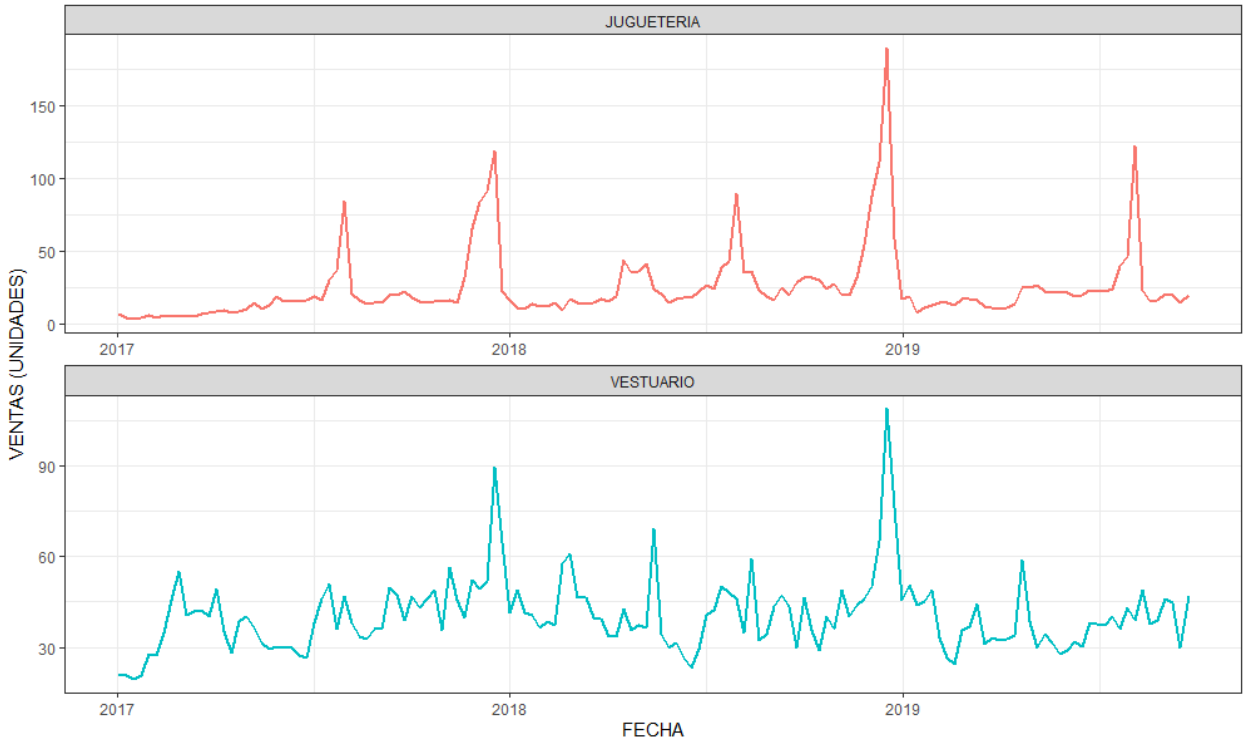

Ilustración 3-1: Promedio de ventas agregadas por sección

Se puede apreciar que para la sección de juguetería existen dos grandes peaks por año, estos son las fechas de navidad y el día del niño, siendo las principales fechas donde se venden estos productos.

En cuanto a la categoría de vestuario, también es posible aprecias grandes ventas a finales del 2018 y 2019, debido a la fecha de navidad. Además de esto, es difícil separar algún efecto de manera clara, esto es debido a que están los productos de esta categoría de manera agregada, y esto puede ocultar los efectos individuales que cada uno presenta. A partir de esto, para visualizar de mejor manera los comportamientos que existen en los productos, a continuación, se muestra un gráfico que muestra algunos de los productos con comportamientos interesantes.

Ilustración 3-2: Ejemplos de comportamientos de productos

En el gráfico 4.2, es posible ver que para el producto A existen alzas de ventas marcadas, esto es debido a que este producto es de juguetería, el cuál mantiene el comportamiento de la sección completa. Para el producto B, se visualiza un comportamiento estacionario, con ventas crecientes en la segunda parte del año, y decrecientes en la primera parte del año. Este comportamiento se da en múltiples series dentro del surtido de productos pertenecientes a vestuario debido a la estacionalidad de los productos, el cual también se puede producir en el sentido contrario, este efecto nos sugiere utilizar modelos de naturaleza estacional.

Dentro del surtido también se analizan productos sin un comportamiento claro, el cual es el caso del producto C del gráfico 4-2. uno de los puntos importantes a considerar en los datos, es que al

ser datos reales del retail, existe bastante ruido en algunos productos, o comportamientos atípicos producidos por razones como quiebres de stock, promociones, u otros eventos del día a día que perturban el comportamiento regular de los productos, esto se debe tener en cuenta al momento del análisis final. Por último, se puede ver que el producto D tiene una baja diferencia entre su punto más alto de ventas y el más bajo, siendo un producto que mantiene unja venta relativamente constante durante el año.

Por lo tanto, tenemos series que tienen un comportamiento con demanda estacionaria, sin tendencia a través del tiempo, con demanda estacional, con peaks según eventos especiales, con comportamiento aleatorio, entre otros. Esto representa un problema para el retail,el cual necesita generar un pronóstico aceptable para cada uno de los productos, existiendo modelos que entienden mejor algunos comportamientos que otros, siendo muy complejo encontrar un único modelo tenga mayor precisión en el pronóstico para todas las series de tiempo, lo que motiva al desarrollo de esta investigación para probar en primer lugar múltiples modelos predictivos de series de tiempo, y posteriormente un clasificador que sea capaz de escoger cada uno de estos dependiendo de la serie de tiempo que se necesite pronosticar.

3.2 METODOLOGÍA

En un principio, se explora y visualiza la data, para entender la naturaleza de las series con las que se trabaja, además, verificar la existencia valores nulos, negativos, outliers, u otros valores atípicos que dificulten la investigación, lo que permite también entender el comportamiento de los datos. En base a esto, se prepara la data para utilizarla directamente en los modelos de pronóstico seleccionados.

Las 5000 series de tiempo serán pronosticadas por cada uno de los siete modelos de pronóstico, lo que permite evaluar diversos candidatos a mejor pronóstico para cada serie, los cuales consideran diferentes factores, y por lo tanto, diferentes resultados. Dentro de los candidatos, están los modelos: media móvil, Holt, Arima, Stlf, NNetAR, TBATS y un ensamble, estos serán detallados más adelante en el marco conceptual.

Se realizará un pronóstico con cada uno de estos modelos con un horizonte de cuatro semanas. Este pronóstico será evaluado en cada una de las semanas por el criterio MAE y WMAPE, cabe destacar que en este caso, donde el peso del WMAPE está dado por las ventas, ambos criterios llevan a la misma conclusión, solo que el primero nos permite observar la diferencia en unidades promedio, y el segundo nos da una referencia del error porcentual, el cuál está ajustado por las ventas. Además de la evaluación ya descrita, se tomará una métrica que nos permita analizar cuántas veces de las 5000 cada modelo es el que genera la predicción más precisa.

Luego, se implementa un octavo modelo a través de meta-learning, el cuál será un clasificador que al ingresar una serie de tiempo, este la analizará y en base a las características de la serie, el clasificador seleccionará el modelo predictivo que tenga mejor rendimiento con ese tipo de series.

La metodología de este clasificador es utilizar como input de entrenamiento los 7 modelos de pronóstico ya evaluados en un grupo de series de entrenamiento entregando como etiqueta qué modelo fue el más preciso para cada serie de tiempo además de generar atributos a partir de cada una de las series. Con esta información, es posible construir un clasificador que nos permitirá decidir qué modelo debemos utilizar para una nueva serie de tiempo en base a características extraídas de ella, sin necesidad de correr todo nuestro abanico de modelos. En este caso el clasificador a utilizar será un random forest.

Por último, se evalúa la precisión del clasificador en su predicción, comparando por MAE/WMAPE y cantidad de veces que fue el modelo más preciso, contra el resto de los modelos predictivos, luego de esto, se generarán conclusiones a partir de todos los modelos, desprendiendo futuros puntos de mejora y concluyendo una recomendación de modelo a utilizar.

3.3 MARCO CONCEPTUAL

Para analizar los resultados, es necesario entender los modelos predictivos a utilizar para abordar las series de tiempo, los modelos de clasificación a utilizar, además de esto se definirán ciertas métricas que ayuden a evaluar el rendimiento de los modelos.

3.3.1 Modelos predictivos

Dentro de los modelos a utilizar se considerarán: Media móvil, Holt, SARIMA, STLF, TBATS, NNetAR y un ensamble. Estos modelos fueron escogidos entre una parrilla de modelos intentando escoger aquellos que capturan distintos comportamientos, dentro de los cuales podemos nombrar características auto regresivas, tendencia, estacionalidad, suavizamientos, funciones trigonométricas y transformaciones de Box-Cox. Además, para visualizar los efectos que captura cada modelo y que estos se entiendan en mayor detalle, se muestra una gráfica con un pronóstico a 1 año para el producto de poleras manga corta con estacionalidad mostrado anteriormente.

3.3.1.1 Media móvil

Una media móvil se define como una secuencia la cual proviene de la media aritmética de n elementos anteriores, en este caso, las ventas obtenidas en los n períodos anteriores con n por definir dependiendo del comportamiento del producto. Esta técnica es útil para tener un primer acercamiento con los datos, y funciona bien al predecir un periodo hacia adelante, para más periodos empieza a tener mayores errores. Además, la predicción de esta se comporta bastante bien al momento de predecir productos con poca variabilidad en sus ventas, por el contrario, falla en los productos de mayor variación, de comportamiento estacional o con tendencia ya que captura estos efectos de forma retrasada.

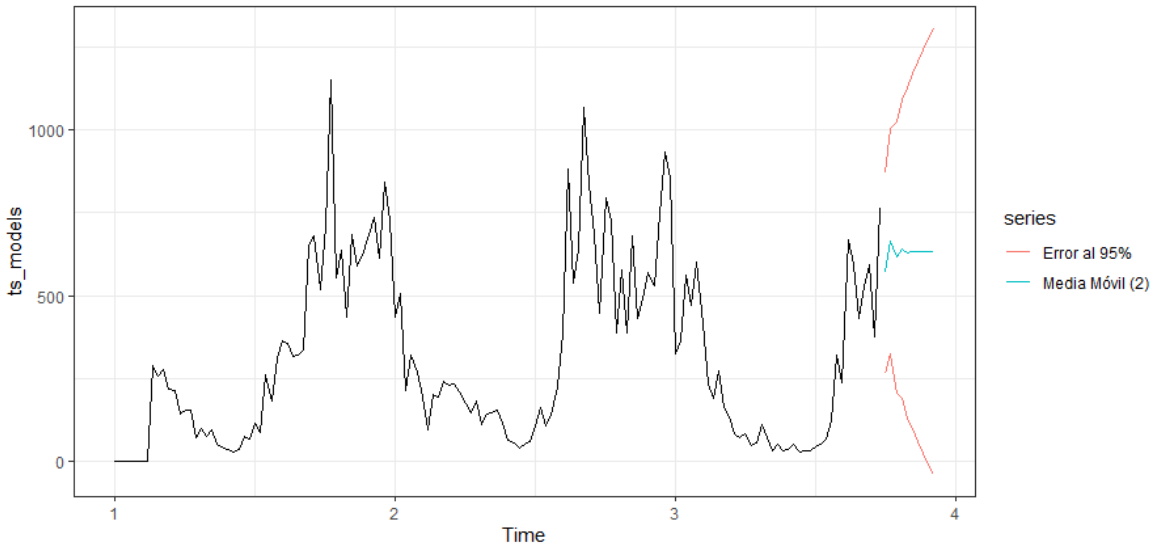

Ilustración 3-3: Modelo Media móvil

3.3.1.2 Holt

El modelo de Holt extiende el suavizamiento exponencial simple e incluye la variable de la tendencia. En particular en este caso se utilizarla variación "damped"/amortiguada, para que este suavice las ventas a largo plazo y así no desviarse en el futuro lejano.

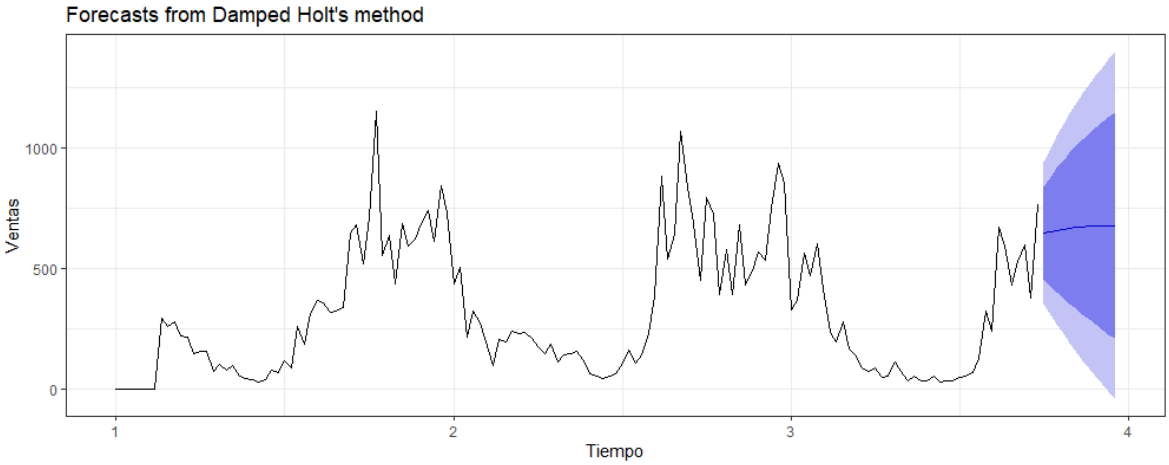

Ilustración 3-4: Modelo de Holt

Se puede ver una línea recta que no hace tanto sentido a primera vista, pero recordar que solo se evaluarán los próximos 4 periodos.

3.3.1.3 SARIMA

El modelo autorregresivo integrado de promedio móvil o ARIMA es un modelo que estudia predicciones de series de tiempo, donde estas series de tiempo pueden considerarse como la realización de un proceso estocástico que se observa secuencialmente a lo largo del tiempo. El modelo ARIMA es un caso particular del modelo ARMA ya que en ARIMA sí existe una raíz unitaria. El modelo ARMA es a su vez es una combinación del proceso autorregresivo AR(p) y el proceso de media móvil MA(q). Ambos procesos son procesos de series de tiempo que intentan explicar la demanda a partir de datos pasados. La diferencia es que el primero tiene memoria a largo plazo por lo que le cuesta reaccionar rápidamente ante perturbaciones (grandes cambios de un periodo a otro), y el segundo tiene corta memoria reaccionando ágilmente a perturbaciones, pero "olvidando" la información del pasado (Brooks, 2008). Este modelo es un buen modelo usado en estadísticas, econometría e ingeniería por varias razones: (i) es considerado como uno de los modelos con mejor desempeño en términos de pronóstico debido a su comprensión de la forma de la serie de tiempo del producto a analizar, (ii) se utilizan como referencia para modelos más sofisticados y (iii) son de fácil implementación y alta flexibilidad dado a su estructura multiplicativa (Dellino, Laudadio, Mari, Mastronardi, & Meloni, 2015). Los parámetros de un modelo ARIMA(p,d,q) se definen como sigue:

- p es el número de términos autorregresivos.
- d es el número de diferencias que se aplican a la serie de tiempo para que sea estacionaria.
- q es el número de medias móviles que realiza el proceso.

En particular, en este caso se llama SARIMA, porque además de los términos ya mencionados, se agregan estos términos también para la estacionalidad, es decir, los parámetros son SARIMA(p, d, q)(P, D, Q)[52].

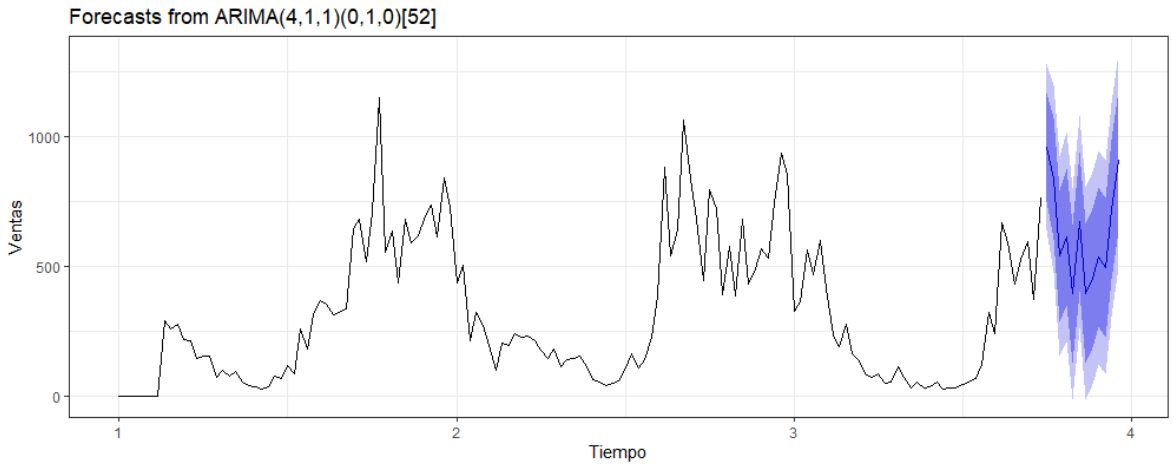

Ilustración 3-5: Modelo ARIMA

Se puede apreciar que es capaz de entender comportamientos estacionales además de cambios abruptos.

3.3.1.4 STLF

El modelo Short tearm load forecasting, es un modelo semi-parametrico aditivo propuesto por Rob Hyndman. (Fan & Hyndman, 2012), el cual asume que una serie de tiempo puede ser descompuesta en error, tendencia y estacionalidad, lo que hace más fácil encontrar patrones y luego crear un pronóstico en base a estos. Considera una descomposición STL más un modelo ETS (Error, Trend & Season).

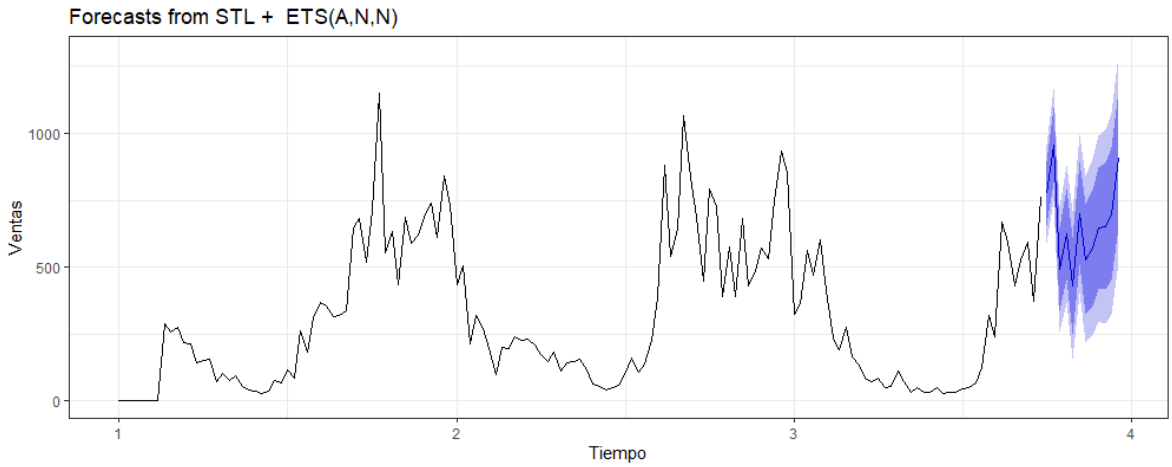

Ilustración 3-6: Modelo STLF

3.3.1.5 TBATS

El modelo TBATS es llamado así por "Trigonometric, Box-cox, ARMA, Trend, Season" , que son todos los parámetros que considera dentro de su cálculo, utilizado para Forecast de modelos de patrones estacionales complejos, atacando ese problema con las funciones trigonométricas, en particular con Fourier (De Livera, Hyndman, & Snyder, 2011).

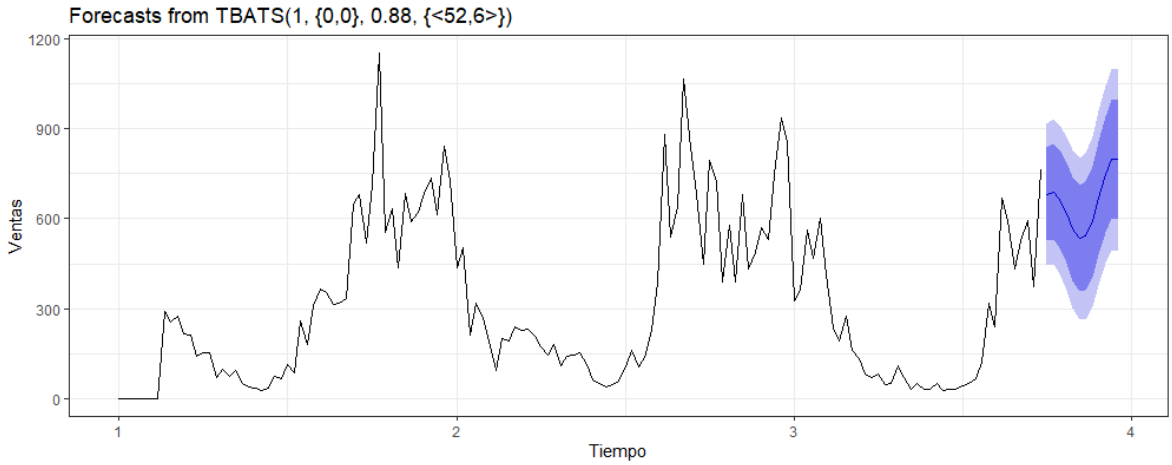

Ilustración 3-7: Modelo TBATS

Es posible apreciar que captura efectos estacionales por el crecimiento en las últimas semanas de pronóstico, pero a su vez se ve un comportamiento suavizado debido a las transformaciones de BoxCox.

3.3.1.6 NNetAR

Redes neuronales son frecuentemente utilizadas para predecir series de tiempo (Dorffner, 1996). Existe una basta cantidad de redes neuronales, en esta oportunidad se usarán las redes neuronales feed-forward error back propagation neural nets, donde las neuronas de la red están organizadas en capas de tal manera que las señales de salida de las neuronas de una capa se transmiten a todas las neuronas de la capa siguiente. En este sentido, el flujo de activación de neuronas va en un solo sentido y pasa capa por capa. El número mínimo de capas que se puede tener son dos capas, la de entrada y la de salida, sin embargo, se pueden agregar capas entremedio llamadas capas ocultas las cuales sirven para aumentar el poder computacional de las redes neuronales, pero a su vez la complejidad del entrenamiento de esta función. Cabe destacar que las redes neuronales se diferencian de los métodos anteriores debido a que estas optimizan y minimizan una función de pérdida, en vez de maximizar la logverosimilitud como es el caso de la regresión lineal múltiple, con un proceso estocástico, donde parto con una solución y actualizo hasta converger, esto puede significar quedar en un óptimo local si el problema no es convexo. El gran beneficio de las redes neuronales en esta oportunidad es que estas son capaces de capturar efectos no lineales, lo que significa información adicional que no siempre es considerada en los modelos anteriores.

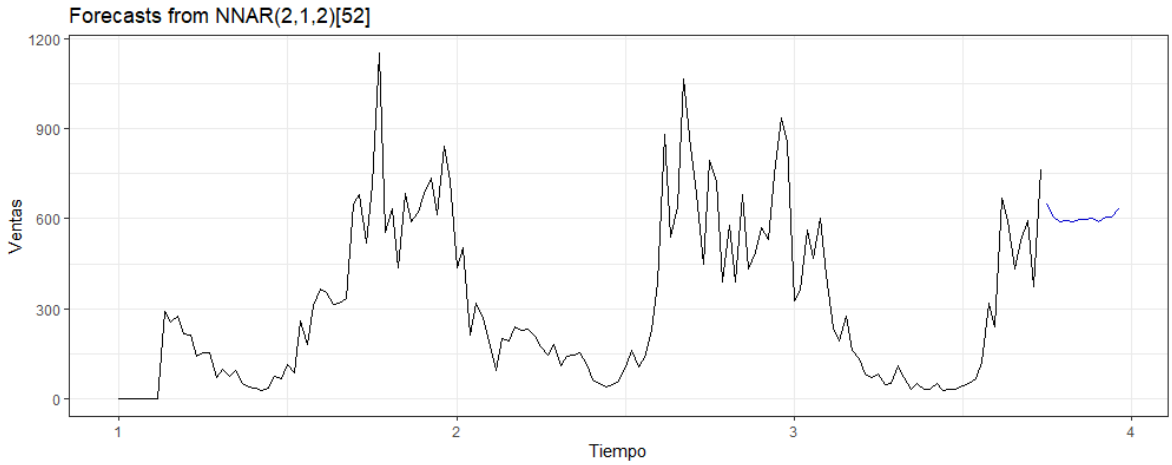

Ilustración 3-8:Modelo NNetAR

3.3.1.7 Ensamble

Un modelo híbrido o ensamble es aquel que busca ponderar dos o más modelos, para así capturar la mayor cantidad de información posible. Para este caso en particular, se buscar generar un ensamble que contenga un modelo que capture los efectos de tendencia que suele funcionar bien en periodos cortos, con un modelo estacional para incorporar ese comportamiento. Existe el riesgo que ambos modelos capturen un efecto y este se amplifique, dando un error mayor a usar solo uno de ellos, pero existe evidencia suficiente para decir que los ensambles en general tienen los mejores rendimientos debido a que un Forecast está basado en variables o información que el otro no considera (Bates & Granger, 1969).

En este caso se utiliza un ensamble entre el modelo de Holt y el modelo TBATS a través de un promedio ponderado simple, esto debido a que ambos observan diferentes comportamientos, el primero enfocado en capturar la tendencia, mientras que el segundo añade el ajuste de la estacionalidad, transformaciones de box-cox, y análisis arma. Existen diversas metodologías para generar ensambles más precisos, pero en esta ocasión en base a resultados preliminares se utilizó esta configuración.

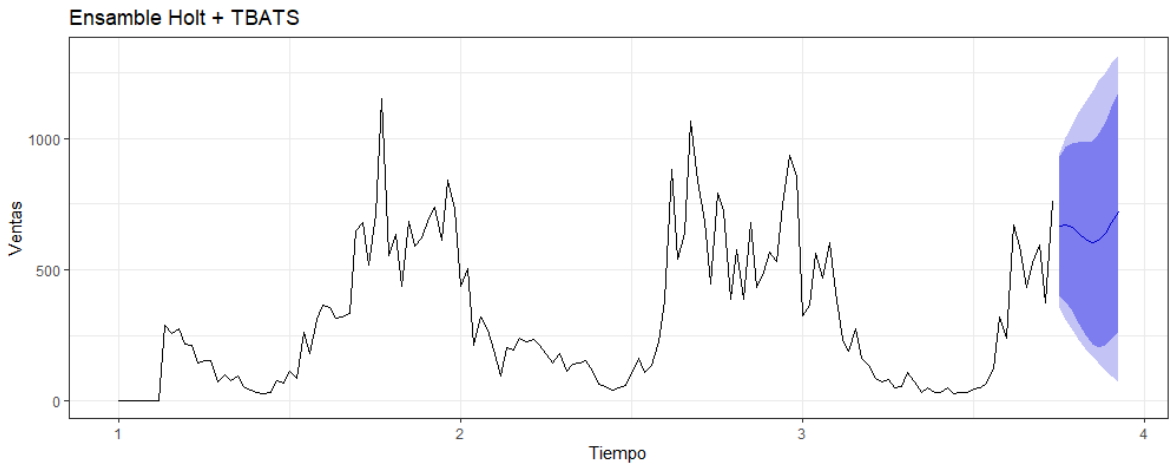

Ilustración 3-9: Modelo ensamblado

Finalmente podemos ver todos los modelos y sus predicciones en un gráfico, para evidenciar la diferencia de comportamiento entre ellos.

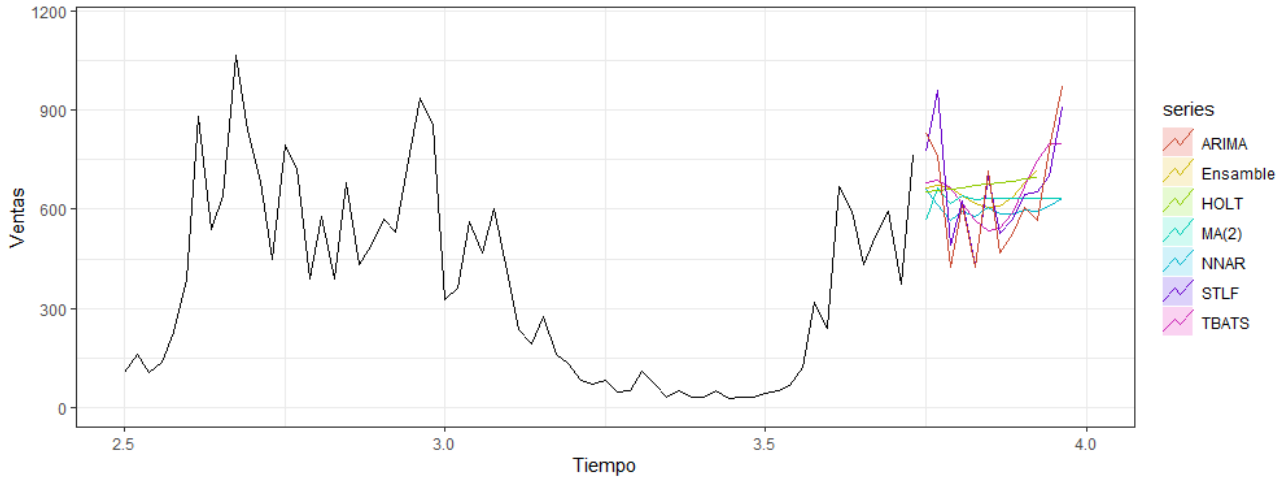

Ilustración 3-10: Pronóstico de todos los modelos utilizados

3.3.2 Modelo de clasificación

Dado que el objetivo de esta investigación es generar la predicción de ventas más certera, su foco está en el pronóstico de series de tiempo, a través de la metodología descrita anteriormente, es posible utilizar un clasificador que decida qué modelo de pronóstico utilizar para cada serie. Para esta investigación se selecciona un modelo de clasificación ya probado con anterioridad, dando espacio para probar otros modelos de clasificación en futuras investigaciones, siendo una de las

posibles variables a ajustar. En este caso se utilizará el modelo de clasificación denominado random forest.

3.3.2.1 Random forest

Un modelo clasificador random forest (Brieman, 2001), es un método de ensamblado que combina un largo número de árboles de decisión (otro método de clasificación) utilizando un método aleatorio a dos pasos, asegurando un clasificador y resultado mejor que cualquier árbol de decisión por sí solo, haciendo el clasificador final más robusto al combinar mil árboles en esta investigación. En este caso, el input será un vector que contenga todas las características de una serie de tiempo, mientras que el output del modelo será que modelo predictivo se debe usar para esa serie de tiempo en específico. En cada nodo de los árboles, será evaluada una característica en base a la etiqueta ganadora entregada, lo que permitirá, luego de ser entrenado, generar una respuesta rápida al entregar un nuevo conjunto de características de una serie. Los resultados de este varían en función al tamaño y calidad del set de entrenamiento, incluyendo la dimensionalidad de este, lo que también influye en su tiempo de ejecución. El resultado final que entregará este modelo será en base a la clase que decidan mayoritariamente los mil árboles, es decir, si 501 árboles optan por el modelo de Holt, esta será la respuesta entregada. En este caso, se utiliza el paquete "randomForest" para la implementación del modelo.

3.3.3 Métricas

Para evaluar la predicción utilizaremos el MAE y el wMAPE. La primera nos servirá para ver el error en unidades, además de ver serie a serie cuál modelo fue más preciso en su predicción (AE), mientras que el wMAPE nos permite evaluar el rendimiento de un modelo para todas las series a nivel porcentual, pero ponderándolo por las ventas de ese producto, dándole mayor relevancia a los productos que tienen más ventas.

3.3.3.1 MAE (Mean Absolute Error)

Es el valor absoluto promedio del error, el cual es definido como la diferencia entre el pronóstico y el valor real.

$$\mathrm{MAE}_t = \frac{\sum_{t=1}^{n} |y_t - f_t|}{N} \quad (1)$$

Siendo y_t = El valor real en el periodo t, y f_t = el valor predicho para el periodo t.

Algunas consideraciones al utilizar el MAE son que este arroja un número en las mismas unidades que la variable de salida, es decir, la diferencia absoluta en ventas, por lo que es fácil de interpretar cuando trabajamos con un producto. Esta métrica depende de la magnitud de las ventas del

producto que se está evaluando, ya que no es lo mismo un MAE de 10 en un producto que vende en promedio 20 unidades, que un MAE de 10 en un producto que vende en promedio 1000 unidades.

3.3.3.2 wMAPE

Weighted Mean Absolute Percentage Error. Esta métrica asigna una ponderación al MAPE en este caso según la cantidad de ventas del producto. Es utilizada para ajustar los casos de ventas bajas, ya que el MAPE se indefine cuando la venta es 0, en cambio, para el wMAPE no influyen las ventas 0, ya que su ponderación es 0 y no es considerado ese valor.

Su valor es el promedio de los errores absolutos multiplicado por las ventas reales y dividido por el promedio de las ventas reales entre todos los productos, en este caso en particular, el peso será dado por el nivel de ventas, por lo que se define de la siguiente manera.

$$\text{wMAPE} = \sum_{i=1}^{n} \left| \frac{y_i - f_i}{y_i} \right| * \frac{y_i}{\sum y_i} = \frac{\frac{1}{n} * \sum |y_i - f_i|}{\frac{1}{n} * \sum y_i} = \frac{MAE}{mean(y)} \qquad (2)$$

Siendo y_i = El valor real de la serie i, y f_i = el valor predicho para la serie i. Se ve que en esta investigación, dónde el peso está dado según la venta, su valor final es el MAE partido en el promedio de las ventas, por lo que estas dos métricas siempre tendrán resultados consistentes entre sí.

3.3.3.3 Modelo más preciso (Win Rate)

Se utilizará una métrica que indica cuál es el modelo más preciso, si esta es sumada a través de las series de tiempo se obtendrá un "win rate", el cuál es definido como el porcentaje de veces que el modelo predictivo es el más preciso versus los demás modelos en el total de las series de tiempo, es decir, si el modelo de Holt tiene la predicción más precisa en 1000 de las 5000 series de tiempo, tendrá un win_rate de 0.2, o 20%. Esta métrica es útil para el desarrollo de esta memoria debido a su rol dentro del meta-learning, donde se necesita diferenciar cual es el mejor modelo para cada serie de tiempo para entrenar el clasificador.

3.4 IMPLEMENTACIÓN Y RESULTADOS DE MODELOS PREDICTIVOS

Cada una de las 5000 series de tiempo será separada en set de entrenamiento y testeo, para así luego poder validar cuál de los modelos es el más preciso. Como se mencionó anteriormente, en base a las necesidades del centro de distribución, es necesario pronosticar cuatro semanas hacia adelante,

dejando el set de entrenamiento desde el “2017-01-02” hasta el “2019-08-26”, mientras que el set de prueba o testeo considera las semanas, 2019-09-02, 2019-09-09, 2019-09-16 y 2019-09-23, lo que nos da horizontes de evaluación de pronosticós desde una hasta cuatro semanas hacia adelante. Una vez transformado cada producto a serie de tiempo, se ingresa a cada uno de los modelos de predicción y se evalúa su desempeño. En las siguientes secciones se utilizará el criterio MAE, pero en las tablas también se encontrará el criterio WMAPE, ambos son consistentes por lo que el análisis es indiferente a cuál de estos criterios se utiliza.

3.4.1 Análisis por criterio MAE

A continuación, podemos ver la evolución del MAE a través de las semanas para cada uno de los modelos para las 5000 series de tiempo estudiadas.

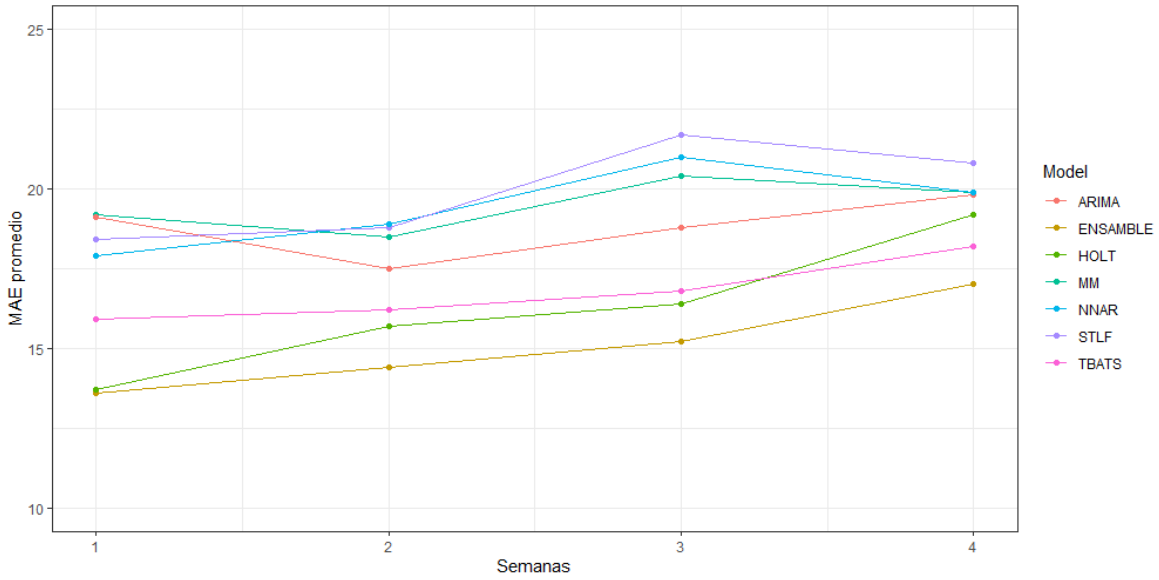

Ilustración 3-11: MAE por modelo a distintos horizontes.

Se puede ver que el modelo que se mantiene con menor MAE a través de las semanas es el modelo de ensamblado, el cual está construido como un promedio ponderado simple entre el modelo Holt y TBATS, a continuación, se muestran tablas detalladas de cada una de las semanas evaluadas, donde es posible analizar otros indicadores. A continuación, se muestran detalladamente los errores de cada modelo para cada semana.

Forecast Semana 1 – 2019-09-02 | Promedio de ventas = 41.2						
MODEL	MAE	Median	Sd	Max	Mae95	wMape
ENSAMBLE	13.6	8	21.2	309	10	33.0%
HOLT	13.7	8	20.5	310	10.2	33.2%
TBATS	15.9	9	29.3	416	12.7	38.6%
NNAR	17.9	11	25	329	13.8	43.5%
STLF	18.4	10	27.3	415	13.6	44.5%
ARIMA	19.1	9	35.4	477	12.8	46.4%
MM	19.2	12	25.5	324	14.8	46.5%
PROMEDIO	16.8	9.57	26	368.57	12.56	40.8%

Tabla 3-1: Evaluación Forecast a una semana

Es posible apreciar que el modelo más preciso según el criterio (Brieman, 2001) MAE es el ensamble, el cual está muy cercano al modelo Holt. Mientras que el modelo con peor rendimiento para la primera semana es la media móvil.

Forecast Semana 2 – 2019-09-09 | Promedio de ventas = 40.1						
MODEL	MAE	Median	Sd	Max	Mae95	wMape
ENSAMBLE	14.4	9	20.2	385	11.1	35.9%
HOLT	15.7	10	21.6	410	12.2	39.2%
TBATS	16.2	10	21.7	391	12.7	39.7%
ARIMA	17.5	10	24.1	454	13.4	40.5%
MM	18.5	12	23	327	14.6	43.6%
STLF	18.8	12	24.3	293	14.6	46.8%
NNAR	18.9	11	28.3	361	14	47.2%
MEAN	17.14	10.57	23	374.43	13.23	41.8%

Tabla 3-2: Evaluación Forecast a dos semanas.

Para el pronóstico con horizonte a dos semanas, existe variación en el rendimiento de los modelos, siendo interesante que tanto la media móvil como el modelo ARIMA, ahora tienen mejor rendimiento que los modelos STLF y NNAR. También se puede ver que la diferencia entre el mejor modelo y el segundo mejor aumentó significativamente, 0.1 unidades de diferencia en promedio, a 1.3 unidades, o desde un 0.2% a un 3.3% de mejora.

Forecast Semana 3 – 2019-09-16 | Promedio de ventas = 26.9						
MODEL	MAE	Median	Sd	Max	Mae95	wMape
ENSAMBLE	15.2	10	18.9	221	11.9	56.2%
HOLT	16.4	11	19.5	220	13	60.7%
TBATS	16.8	10	22.7	314	13	62.4%
ARIMA	18.8	11	26.1	359	14.2	69.5%
MM	20.4	13	25.9	404	16	75.7%
NNAR	21	12	30.1	642	15.9	77.9%
STLF	21.7	12	31.1	418	16.3	80.4%
MEAN	18.61	11.29	25	368.29	14.33	69.0%

Tabla 3-3: Evaluación Forecast a tres semanas

Para el pronóstico a tres semanas, se observa que la diferencia entre los dos primeros modelos se mantiene. Además, se ve un incremento considerable en el WMAPE, de un promedio de 40% a un 69%, esto se debe a que el promedio de ventas bajó la semana del 16 de septiembre, ya que esa fue la semana en donde se celebran las fiestas patrias en Chile, en particular, el año 2019 tuvo una gran cantidad de días festivos y las ventas se vieron afectadas directamente por tener que cerrar las puertas de los locales. Además, podemos ver que el MAE promedio subió a pesar de que se vendieron muchas unidades menos, es decir, el WMAPE alto nos evidencia que ningún modelo pudo predecir esta baja en las ventas, lo que tiene sentido ya que años anteriores no existieron tantos días festivos para las fiestas patrias, por lo que las ventas no se vieron afectadas como el 2019.

Forecast Semana 4 – 2019-09-23 | Promedio de ventas = 41.7						
MODEL	MAE	Median	Sd	Max	Mae95	wMape
ENSAMBLE	17	10	24.5	484	12.9	40.8%
TBATS	18.2	11	25.7	474	13.9	43.6%
HOLT	19.2	11	26.8	494	14.7	46.0%
ARIMA	19.8	10	31.7	601	14.4	47.5%
MM	19.9	12	27.5	384	15.3	47.6%
NNAR	19.9	11	27.3	340	15.1	47.7%
STLF	20.8	11	30.2	459	15.7	49.9%
MEAN	19.26	10.86	28	462.29	14.57	46.2%

Tabla 3-4: Evaluación Forecast a cuatro semanas

El pronóstico a cuatro semanas mantiene la tendencia que se veía anteriormente, algo relevante es que el modelo TBATS fue el que menor incremento su error a través de las semanas, quedando como el segundo mejor modelo si se mira en un horizonte a cuatro semanas según criterio MAE.

Resumen MAE a través de las semanas					
Modelo	S1	S2	S3	S4	Promedio por modelo
ENSAMBLE	13.6	14.4	15.2	17	15.1
HOLT	13.7	15.7	16.4	19.2	16.3
TBATS	15.9	16.2	16.8	18.2	16.8
NNAR	17.9	18.9	21.7	19.9	19.6
STLF	18.4	18.8	21	20.8	19.8
ARIMA	19.1	17.5	18.8	19.8	18.8
MM	19.2	18.5	20.4	19.9	19.5
Promedio por semana	16.8	17.1	18.6	19.3	

Tabla 3-5: MAE modelos según horizonte de predicción

Se ve que el MAE promedio entre los modelos va aumentando a medida que el pronóstico se realiza a un mayor horizonte, esto tiene sentido, debido a que existe mayor incertidumbre en el futuro lejano que en el cercano. Podemos ver que al modelo de Holt fue al que más le afectó esto, aumentando su error considerablemente en las últimas semanas con respecto a las primeras.

El modelo con mayor precisión a través de las semanas fue el ensamble construido a partir del modelo TBATS y el modelo de Holt, esto se puede deber como fue dicho anteriormente, a que considera información de más de un modelo.

3.4.2 Análisis por mejor predicción

Luego de tener el MAE de cada modelo para cada serie de tiempo, se puede verificar cuál es el modelo con una predicción más precisa (menor error) para cada serie de tiempo, y en que porcentaje de las 5000 series es el mejor. Cabe mencionar que si dos modelos son igual de precisos (ejemplo, ambos tienen MAE = 1), ambos son modelos ganadores, por lo que los porcentajes pueden dar más de 100%, debido a que la única forma de lograr este, es que solo un modelo fuera el ganador para cada serie. A continuación, vemos para cada semana cuánto "gano" cada modelo.

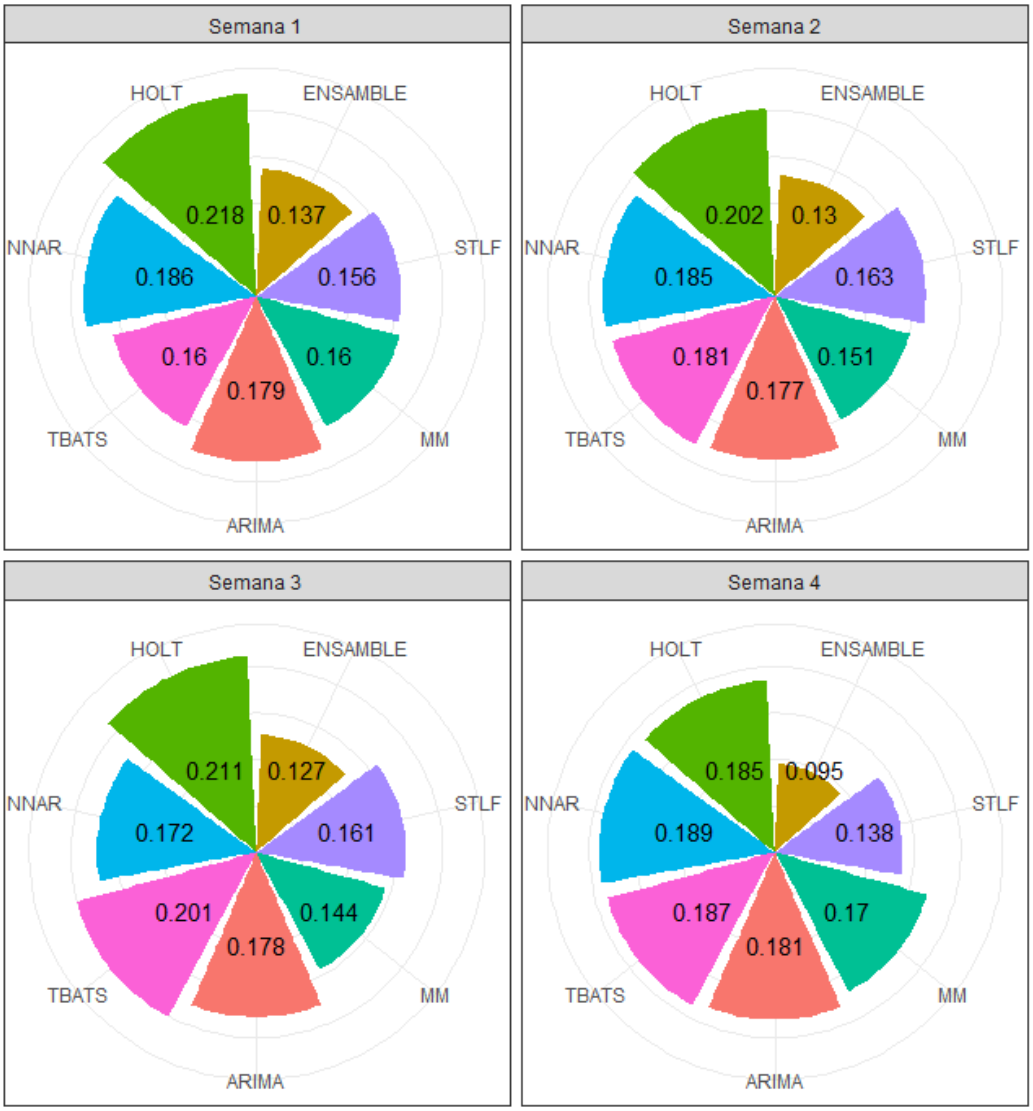

Ilustración 3-12: Win rate de cada modelo según MAE para cada horizonte de pronóstico

Como se puede ver en la ilustración, ningún modelo es completamente dominante sobre otros, ya que el más preciso solo tiene una mejor performance en el 21.8% de los casos, que es el modelo Holt para la primera semana de pronóstico.

A continuación, se muestra una tabla resumen del rendimiento de cada uno de los modelos en los distintos horizontes según porcentaje de veces que obtuvieron la mejor predicción para las 5000 series de tiempo.

	HOLT	NNAR	ARIMA	TBATS	MM	STLF	ENSAMBLE
Semana1	**21.8%**	**18.6%**	17.9%	16.0%	16.0%	15.6%	13.7%
Semana2	**20.2%**	**18.5%**	17.7%	18.1%	15.1%	16.3%	13.0%
Semana3	**21.1%**	17.2%	17.8%	**20.1%**	14.4%	16.1%	12.7%
Semana4	18.5%	**18.9%**	18.1%	**18.7%**	17.0%	13.8%	9.5%

Tabla 3-6: Win rate de modelos según MAE para cada horizonte de pronóstico

Por otro lado, es interesante darse cuenta de que si bien el modelo NNAR es el menos preciso según el criterio MAE para la segunda semana (Tabla MAE 2s), es el modelo con segundo mayor win rate cuando se trata de combinaciones en las que fue el más preciso, lo cual concuerda con la alta desviación estándar presentada por este modelo. Esto nos quiere decir que si bien, en promedio el modelo NNAR no es el mejor, es bastante preciso en ciertos casos, mientras en otros su predicción es de las más alejadas a la venta real. Dado este análisis, se vuelve interesante la idea de poder aprovechar este modelo en los casos donde tiene mejor rendimiento, y escoger otro modelo cuando este no es capaz de hacer una buena predicción.

3.5 CONCLUSIONES MODELOS PREDICTIVOS

Los resultados de las predicciones de los distintos modelos indican que el modelo más estable a través de las semanas es el ensamble, siendo en promedio 4.4 unidades más certero que la media móvil, el cuál es el objetivo por superar. Además de esto, el WMAPE de la media móvil promedio a través de las semanas fue de un 53.4%, mientras que del ensamble fue 41.5%, es decir, el ensamble es un 11.9% más certero.

Evaluando en la última métrica considerada, la media móvil es uno de los modelos que tiene la mejor predicción para la menor cantidad de series según el criterio de win_rate, obteniendo un 15.6% a través de las semanas, superado ampliamente por el modelo de Holt que fue el mejor un 20.4% de las veces, es decir, una diferencia de un 4.8%. También es importante destacar que esta métrica no es necesariamente consistente con el MAE, podemos identificar que a pesar de que el modelo Ensamblado es el que tiene menor MAE en el pronóstico a cuatro semanas, es el que gana en menor cantidad de series de tiempo según el criterio win_rate, por eso ambas métricas son necesarias y se debe tener en consideración la diferencia entre ambas.

Es justo destacar que el ensamble no solo superó ampliamente a la media móvil, si no que a todos los demás modelos según el criterio MAE, por lo que es recomendable considerarlo en los distintos análisis a futuro.

Otro resultado relevante en el punto 4.4.2, es que no existe un modelo con la mejor predicción para todas las series, siendo el modelo de Holt el que mejor logra esto, alcanzando la predicción más certera en un 21.8% por ciento de las 5000 series de tiempo para la primera semana, esto da espacio para pensar que una estrategia fuerte sería considerar distintos modelos dependiendo del tipo de serie de tiempo el cuál se quiere predecir, esto será explorado en profundidad en la siguiente sección.

4 PREDICCIÓN A TRAVÉS DE META-LEARNING

Dada la evidencia entregada en el punto anterior, es posible pensar que, si existiera una forma de describir una serie de tiempo, y en base a esta descripción se pudiera saber de antemano qué modelo tendría un mejor rendimiento para esa serie en particular, sería posible ocupar el modelo adecuado para cada serie de tiempo, lo que permitiría tener una predicción mejor en la mayoría de los casos por sobre usar cada modelo por si solo para todas las series. Es decir, el objetivo es crear un modelo de clasificación que reciba como input una serie de tiempo luego calcule características de esta, y finalmente entregue el modelo de pronóstico a utilizar. Para esto es necesario entrenar este clasificador, entregándole un set de series con características, y cuál fue el mejor modelo dadas esas características. Ya tenemos el mejor modelo para cada serie del punto anterior, por lo que ahora se calcularán atributos para las distintas series de tiempo.

4.1 METODOLOGÍA

Generar base con todas las series de tiempo disponibles, y con los atributos de cada una de estas, además de una etiqueta que indique cuál fue el modelo con una predicción más certera para esta serie.

Luego, se separa la base en entrenamiento y testeo, en este caso se utilizará 80% entrenamiento y 20% testeo. Se entrena un modelo clasificador random forest, con la base de entrenamiento.

Se predice que modelo utilizar para el 20% de las series perteneciente al testeo en base a sus atributos, y luego se evalúa que hubiese pasado si se ocupaba el modelo que el clasificador decía, esto se evalúa tanto en MAE, como en cuántas veces hubiese sido el modelo más preciso.

Para entender en qué casos funciona esta metodología, se probará esto con cada una de las 4 semanas, además primero se probará introduciendo solo los modelos de peor rendimiento, para ver si el MAE mejora, luego los de mejor rendimiento, luego todos los modelos, y finalmente se probará contra todos los modelos, pero entrenado solo con los mejores modelos.

Es importante recordar que el objetivo de la investigación es obtener la mejor predicción según criterio MAE, por lo que una mejora de MAE por parte del clasificador frente a cada modelo por si solo es relevante.

4.2 GENERACIÓN DE ATRIBUTOS PARA SERIES DE TIEMPO

Para generar atributos/características para cada serie de tiempo, se utilizará el paquete “tsfeatures” desarrollado por Rob J Hyndman y Yangzhuoran Yang, el cual contiene funciones que, al entregar una serie de tiempo en cierto formato, calculan con los valores de la serie distintos atributos de esta. En particular en esta investigación se utiliza un set de 17 atributos, entre los que se encuentran, la tendencia, la fuerza de su estacionalidad, la entropía, su estabilidad, la linealidad de la serie, la

curvatura, entre otras, entregadas por la función “tsfeatures”. Para ver el set de atributos y su construcción en detalle es posible visitar la librería tsfeatures en CRAN.

A modo de ejemplo, a continuación, se muestra una de las características, la “fuerza de estacionalidad”, la cual va entre 0 y 1, y mide que tan marcada es la estacionalidad en una serie. En la ilustración 5-1 se muestra una alta fuerza estacional de 0.931 y se puede apreciar ese efecto visualmente, mientras que la ilustración 5-2 tiene una baja fuerza estacional de 0.254, y no es posible apreciar algún patrón fácilmente.

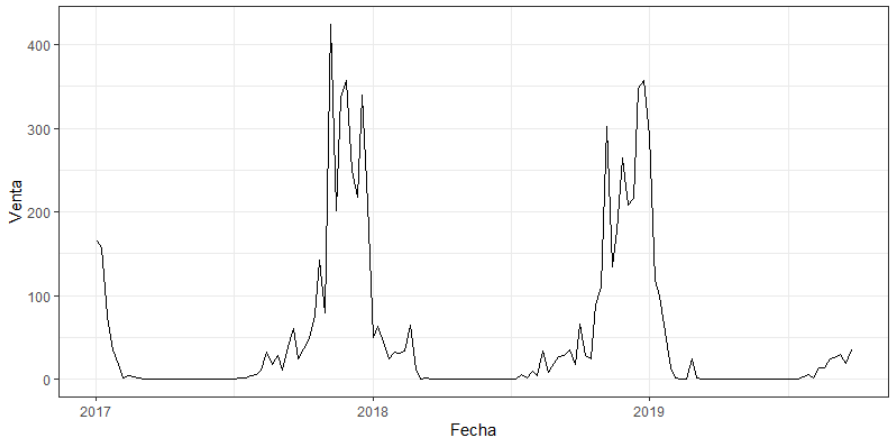

Ilustración 4-1: Serie de tiempo Seasonal Strenght = 0.931

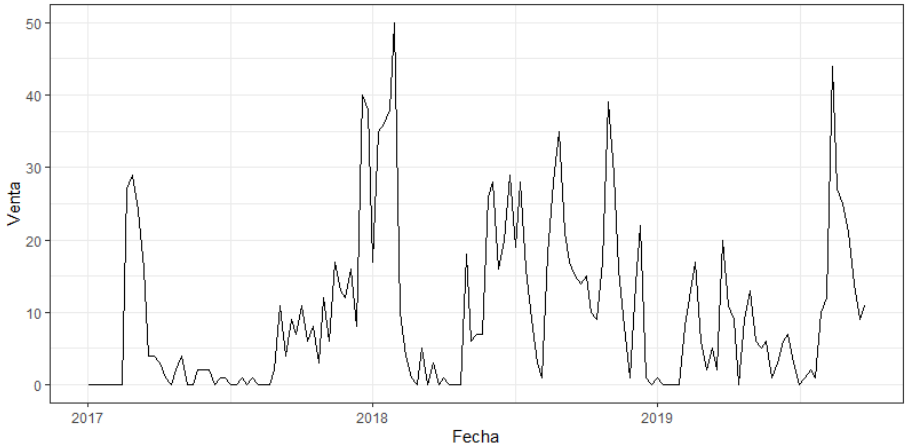

Ilustración 4-2: Serie de tiempo Seasonal Strenght = 0.254

Es importante entender que estos atributos permiten clasificar nuestras series de tiempo, para así, describir cada una de estas y luego con estas características poder decidir qué modelo predictivo es el más adecuado en base a esa descripción y serie de tiempo.

Se visualizan algunos atributos adicionales en el siguiente gráfico, donde cada fila/columna es uno de los atributos (trend, spike, linearity, curvature, e_acf1, e_acf10), y cada punto, es una de las 5000 series representada por esa característica.

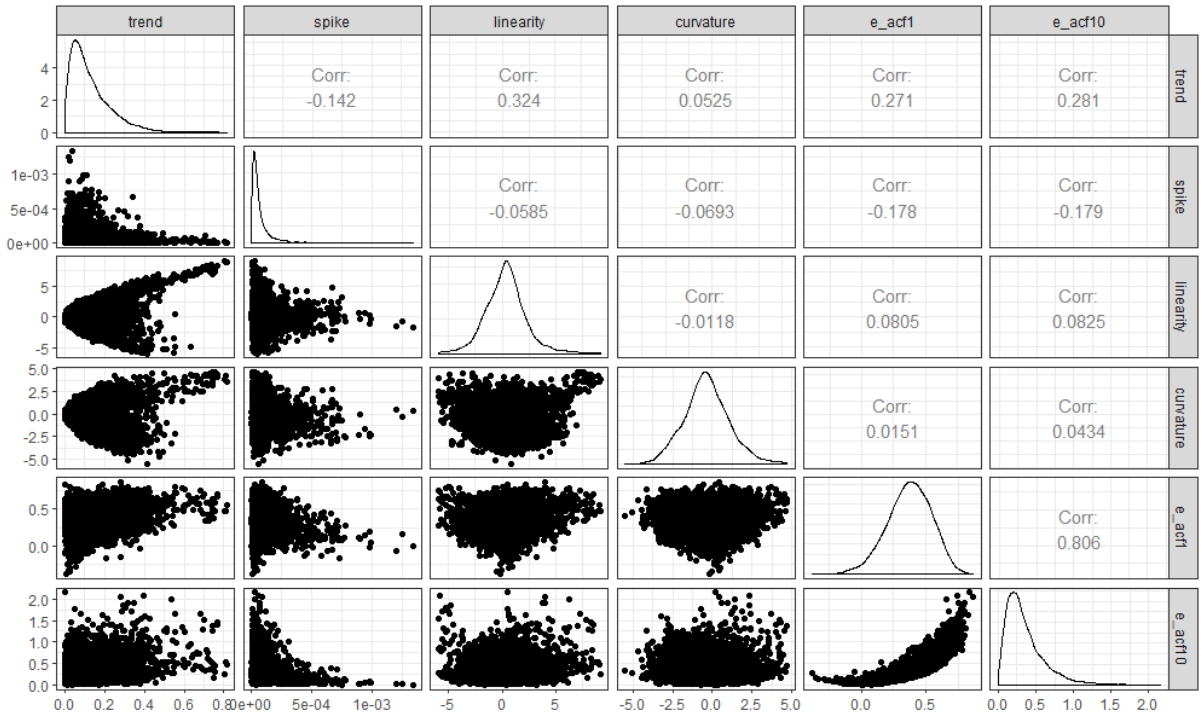

Ilustración 4-3: Relación entre las características de las 5000 series de tiempo

Algo interesante a mostrar, es que los atributos tienen variación en los valores, es decir, efectivamente se identifican series similares, y otras diferentes entre sí, esto nos da indicios de que contamos con un conjunto inicial mínimo lo suficientemente diverso para luego predecir a partir de nuevos valores.

Es relevante que estos atributos sean confiables, ya que de ellos depende un correcto entrenamiento del modelo de clasificación, y de la posterior predicción, y para que estos atributos sean confiables, nuestras series de tiempo deben serlo. Dado que se trabaja con datos reales del retail, existen series que comienzan las ventas más tarde, ya que en el local antes no existía ese producto, lo que distorsiona el cálculo de algunos atributos para algunas series, un ejemplo de esto se ve en la siguiente serie.

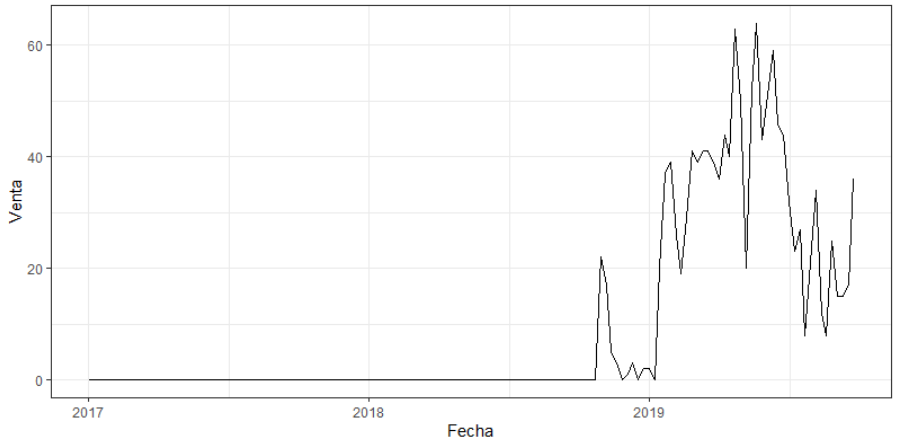

Ilustración 4-4: Serie con trend = 0.767, cálculo alterado por ventas 0

El valor de tendencia para esta serie es de 0.767, uno de los más altos, pero esto no necesariamente está relacionado con que la serie como tal tenga una gran tendencia, sino que, al ser consideradas las transacciones desde el 2017, y las ventas de este producto comenzaron a fines del 2018, es considerada como una tendencia alta. Esto puede generar un error en el entrenamiento del clasificador si es muy frecuente, y es un punto a tener en cuenta.

4.3 ENTRENAMIENTO MODELO RANDOM FOREST

Se utilizará el modelo random forest debido a su capacidad de considerar múltiples árboles de decisión generando un resultado más robusto. En este caso se elige una cantidad de 1000 árboles debido a que los resultados no varían con una cantidad superior y se utilizan los atributos mencionados en la parte anterior 4.2.

Este es entrenado con el 80% de la data de entrenamiento, y luego se procede a predecir al 20% restante. Tal y como se menciona en el punto 4.1, esto se hace 4 veces para la primera semana de predicción, evaluando los resultados dependiendo de que modelos se le ingresen.

Dado que existen series de tiempo que tienen más de un modelo ganador bajo el criterio MAE, debido a que su predicción es igual de certera, se excluirán estos casos del entrenamiento del modelo, ya que entregar dos etiquetas (modelos ganadores) para los mismos atributos (mista serie de tiempo) entorpecería en el entrenamiento del modelo. De las 5000 series de tiempo entregadas, existen 1103 series que tienen un empate en la precisión de su predicción al menos entre dos modelos, por lo que para el clasificador serán utilizadas las 3897 series restantes.

4.3.1 CASO 1: Los tres modelos con mayor MAE

Se entrena el clasificador random forest entregando etiquetas solamente de los modelos con más bajo rendimiento, que bajo el criterio MAE, en el caso de pronóstico a una semana, fueron los modelos Arima, Media móvil y STLF. Al predecir la base de testeo se obtienen los siguientes resultados sobre la precisión del clasificador.

Matriz de confusión				
	ARIMA	**MM**	**STLF**	**class.error**
ARIMA	940	225	198	31.0%
MM	481	309	191	68.5%
STLF	522	265	165	82.7%

Tabla 4-1: Matriz de confusión clasificador a una semana con los modelos de menor rendimiento

A la izquierda de la matriz se muestran las etiquetas reales mientras que en la parte superior las etiquetas predichas, es decir, si vemos la primera fila (ARIMA), primero veremos cuantas veces acertamos en predecir el modelo ARIMA (940), luego, vemos que en 225 casos el modelo predijo media móvil cuando en verdad era ARIMA, y por último, el modelo predijo en 198 casos que el mejor modelo era el STLF cuando en verdad el mejor era el modelo ARIMA, y así con los demás modelos. Además, es posible ver el error de la predicción de cada clase, calculado como el complemento de la precisión, que es cantidad de aciertos/cantidad total de ese modelo, en el caso del modelo ARIMA, 1 -940/1363 = 0.26 = 31% de error.

Se puede ver en la matriz de confusión que en general se acierta la predicción del modelo ARIMA cuando este es el ganador, mientras que para la media móvil y el modelo STL presenta un gran error. Recordando la tabla de MAE a una semana, el modelo arima era el de casi mayor MAE, solo superado por la media móvil en 0.1 unidad, pero al mismo tiempo, entre los tres modelos acá expuestos, era el con mayor win_rate (Tenía la predicción más precisa en más series de tiempo). Ahora se muestra en el set de testeo la comparativa por MAE y ratio de victorias de la predicción del modelo de clasificación “CLAS” con los tres modelos utilizado para entrenar.

MODEL	**MAE**	**WIN_RATE**
CLAS	**16.8**	**43.2%**
STLF	19	28.9%
ARIMA	20.3	41.3%
MM	20.4	29.8%

Tabla 4-2: Comparación entre modelo de clasificación a una semana con modelos de menor rendimiento

Es posible ver que obtiene un MAE menor que cualquier de los modelos por separados, además, tiene la mejor predicción en mayor cantidad de series de tiempo, lo que nos dice que el clasificador elige de manera correcta para modelos que lo hacen bien para algunas series, pero mal para otras.

4.3.2 CASO 2: Dos mejores modelos según MAE

Ahora, se entrena el clasificador random forest entregando etiquetas solamente de los modelos con más alto rendimiento, que bajo el criterio MAE, en el caso de pronóstico a una semana, fueron los modelos de Ensamble y Holt. Al predecir la base de testeo se obtienen los siguientes resultados sobre la precisión del clasificador.

Matriz de confusión			
	ENSAMBLE	**HOLT**	**class.error**
ENSAMBLE	1017	628	**38.2%**
HOLT	882	602	59.4%

Tabla 4-3: Matriz de confusión clasificador a una semana con los modelos de rendimiento más alto

Se ve que la precisión es mayor para el modelo del ensamble, a continuación, se hace la validación en el set de testeo de cuántas veces obtiene la predicción más precisa en comparación a los dos modelos, y cuál sería el MAE en caso de obedecer los resultados entregados por el clasificador.

MODEL	**MAE**	**WIN_RATE**
CLAS	**13.3**	**53.6%**
HOLT	13.5	47.4%
ENSAMBLE	13.6	52.6%

Tabla 4-4: Comparación entre modelo de clasificación a una semana con modelos de mayor rendimiento en set de testeo

Se aprecia que el MAE de los tres modelos es muy similar, al igual que el win_rate, de igual manera en ambos casos el clasificador obtuvo una pequeña ventaja. También es interesante notar que a pesar de que el ensamble tenga una predicción más precisa en más series de tiempo, obtiene un mayor MAE, esto se debe a cuanto más precisa es la predicción.

4.3.3 CASO 3: Todos los modelos según MAE

En este caso, se entrena el clasificador random forest entregando etiquetas de todos los modelos ejecutados, en el mejor de los casos se espera que el clasificador pueda capturar toda esta información y predecir de manera exitosa cuál modelo es más conveniente para cada serie. Al predecir la base de testeo se obtienen los siguientes resultados sobre la precisión del clasificador.

Matriz de confusión								
	ARIMA	**ENSAMBLE**	**HOLT**	**MM**	**NNAR**	**STLF**	**TBATS**	**class.error**
ARIMA	56	15	133	46	90	39	35	86.5%
ENSAMBLE	30	8	101	22	52	35	20	97.0%
HOLT	62	17	257	59	89	54	31	54.8%
MM	41	7	141	67	82	53	25	83.9%
NNAR	60	13	160	58	120	51	24	75.3%
STLF	40	11	140	56	74	64	26	84.4%
TBATS	46	6	133	41	84	52	14	96.3%

Tabla 4-5: Matriz de confusión clasificador a una semana con todos los modelos

En este caso se ve que el error es mucho mayor en promedio para todas las clases, esto es debido a que mientras más clases haya por predecir, en promedio mayor será la posibilidad de equivocarse.

Se puede ver al sumar cada columna que el modelo que prefiere en la mayoría de los casos el clasificador es el de HOLT, siendo en el que tiene mayor precisión, lo cuál es bueno, debido a que este modelo era el más preciso la mayor cantidad de veces según el análisis realizado en la tabla 4-7. Consistentemente, el menos predicho y con mayor error es el ensamble, esto no necesariamente es bueno, ya que, según nuestros análisis anteriores, a pesar de ser uno de los modelos que menor cantidad de veces era el más preciso, en promedio era el que tenía mejores resultados de MAE, criterio por el cuál medimos la precisión de la predicción finalmente, esta información se utilizará para el caso 4.

Dado que a nosotros no nos interesa solo acertar al mejor modelo, sino que también cuál es el MAE de lo dicho por el clasificador y cuántas veces ganó en comparación al resto de los modelos, se analiza la siguiente tabla.

MODEL	**MAE**	**WIN_RATE**
ENSAMBLE	**12.7**	9.2%
HOLT	12.9	**19.3%**
TBATS	14.8	12.8%
CLAS	14.9	**19.3%**
STLF	17	14.0%
NNAR	17.2	16.5%
MM	18.3	14.2%
ARIMA	18.5	14.1%

Tabla 4-6: Comparación entre modelo de clasificación a una semana con todos los modelos

Podemos ver que en cuanto a criterio MAE, los mejores modelos fueron el Ensamble y el de Holt, dejando en cuarto lugar al clasificador construido. Por otro lado, en cuanto a cantidad de veces que tiene la mejor predicción sobre el total de series, el modelo de Holt y el clasificador son los que

tienen mayor win_rate. Esto se puede explicar porque el clasificador es construido en base al criterio de modelo ganador en cada serie, y no sobre el MAE. Lo que nos dice que el clasificador funciona si el objetivo es ser el mejor modelo la mayor cantidad de veces, pero no se asegura aún que tenga el mejor MAE.

4.3.4 CASO 4: Todos los modelos entrenando solo con los dos mejores

En este caso, se utiliza el clasificador entrenado en el CASO 2, tomando como input solo los modelos de Holt y Ensamble, esto es debido a que nosotros de antemano sabemos que estos modelos son los que tienen mejor performance a nivel de MAE/WMAPE, y haremos uso de esa información valiosa, significando que el clasificador solo nos predecirá estos dos modelos a pesar de que lo compararemos con todos. Al predecir la misma base de testeo utilizada en la parte anterior del CASO 3, obtenemos los siguientes resultados.

MODEL	MAE	WMAPE	WIN_RATE
CLAS	**11**	**27.1%**	**24.9%**
ENSAMBLE	12.7	31.3%	9.2%
HOLT	12.9	31.8%	19.3%
TBATS	14.8	36.5%	12.8%
STLF	17	41.9%	14.0%
NNAR	17.2	42.4%	16.5%
MM	18.3	45.1%	14.2%
ARIMA	18.5	45.6%	14.1%

Tabla 4-7:Comparación entre modelo de clasificación a una semana con modelos, entrenado con los mejores dos

Podemos ver que, al entrenarlo solo con los dos mejores modelos, tiene un MAE bastante mejor que al entrenarlo con todos los modelos, esto se puede deber a que, en el caso anterior, al poder entregar más etiquetas como resultado, equivocarse era más costoso, debido a que el error de equivocarse a nivel de MAE podía ser muy grande. Predecir un ARIMA cuando realmente no era el modelo correcto te acerca al 18.5 de MAE promedio que tiene ese modelo por sí solo, mientras que predecir un ensamble y equivocarse, te acerca a un MAE de 12.7, que es uno de los mejores sin contar nuestro clasificador. Además, es importante mencionar que este set de testeo tiene una venta promedio de 40.6 unidades, lo que significa que **el modelo clasificador tiene un wMape de 27.1%, mientras que la media móvil tiene wMape de 45.1%**, **generando una predicción un 18% más precisa**. Estos resultados son muy relevantes, ya que evidencian que existe un amplio espacio de mejora en la predicción de ventas.

4.4 ANÁLISIS DE LOS RESULTADOS

Cuando se entrena al clasificador con pocas etiquetas, este es capaz de mejorar el rendimiento de la predicción satisfactoriamente, tanto a nivel de MAE, como a nivel de win_rate, es decir, en cuántas series de tiempo es el modelo más preciso. Esto se ve reflejado tanto en el experimento con los tres peores modelos, como en el caso de los dos mejores.

Si bien se aprecian errores altos al momento de visualizar la matriz de confusión, los que podrían llevar a concluir que se tiene un "mal clasificador", el costo de equivocarse de etiqueta puede ser relativamente bajo, es decir, si el mejor modelo era ARIMA, pero el clasificador predice ENSAMBLE que estaba solamente una unidad desviado con respecto al mejor modelo, se considera como error, pero no tiene mayor impacto en el MAE al momento de sumar todas las series. Para mejorar la precisión del clasificador es necesario cuestionar los atributos que son calculados de las series de tiempo y la calidad de estas mismas tanto en el entrenamiento como en el testeo.

No se pueden asegurar resultados confiables en cuanto a MAE si es que el input para entrenar el clasificador son modelos muy variados, sobre todo si es que algunos de estos tienen un mal rendimiento, lo que perjudica el rendimiento del clasificador en base a la métrica MAE. En ese caso, es recomendable aprovechar la información obtenida de los experimentos anteriores, e incluir en el entrenamiento del clasificador solo modelos que ya presentan un buen nivel de predicción. Esto se debe a que no es claro que modelo será el mejor en base a los atributos contribuidos, pero si está lo suficientemente cerca para mejorar la precisión versus no utilizar el clasificador.

En base a los 4 casos realizados, y el win_rate asociado a cada modelo, podemos concluir que el clasificador random forest es el que tiene una mejor predicción para la mayor cantidad de series de tiempo, además de tener un mejor error absoluto medio (MAE) para este set de test como se observa en el CASO 4, en donde el clasificador le gana al siguiente modelo (ensamble) en un 4.2% y a la media móvil en un 18%.

5 CONCLUSIONES

A partir de la comparación de los resultados obtenidos por los distintos modelos predictivos implementados a un horizonte de cuatro semanas realizada en el punto 3, se concluye que el modelo predictivo más estable a través de las semanas es el ensamble construido entre el modelo de Holt y el modelo TBATS, siendo este en promedio 4.4 unidades más certero que la media móvil, el cuál es el objetivo por superar. Por otro lado, el WMAPE de la media móvil promedio a través de las semanas fue de un 53.4%, mientras que del ensamble fue 41.5%, es decir, el ensamble es un 11.9% más certero que la media móvil. Esto nos lleva a concluir que aplicando modelos predictivos más sofisticados es posible aumentar la precisión en un 11.9% para este tipo de productos que tienen un comportamiento particular.

Con respecto al clasificador a través de meta-learning desarrollado en el punto 4 se prueba que, de ser entrenado adecuadamente, un modelo clasificador random forest es capaz de elegir de manera aceptable para una serie de tiempo, que modelo predictivo debe implementarse con ella en base a las ventas que ha tenido en los últimos 2 años en este caso, obteniendo un error promedio absoluto menor a cualquier modelo por si solo en un set de múltiples series de tiempo (1000 presentes en el testeo en los diferentes casos del punto 4.3). Es necesario aclarar que este debe ser entrenado de manera inteligente, aprovechando toda la información disponible, en este caso por ejemplo, el mejor clasificador fue el entrenado solamente con dos de los siete modelos posibles, debido a que entrenarlo para predecir siete modelos, terminaba siendo contraproducente, y aumentando el error promedio absoluto al evaluarlo con el set de testeo, obteniendo un rendimiento peor que otros tres modelos individuales en ese caso en particular (ver 4.3.3 - CASO 3).

Dado que la metodología para crear el modelo clasificador a través de meta-learning fue exitosa en términos de MAE, WMAPE y win_rate, se puede concluir que esta sirve para el pronóstico de múltiples series de tiempo relacionadas con productos del retail, pero se deben tener precauciones, debido a que una mala implementación puede llevar a resultados deficientes.

Es posible, y se recomienda implementar un modelo más sofisticado que la media móvil, debido a que la diferencia en la precisión del pronóstico de la media móvil con el resto de los modelos predictivos es de una magnitud considerable, llegando a **un máximo de hasta un 18% de diferencia en la precisión entre el mejor modelo implementado y la media móvil**, según el experimento realizado en el punto 4.3.4 CASO 4, siendo una diferencia mayor que la obtenida por el ensamble contra la media móvil, la cual era de un 11.9%. De todas formas, la elección del modelo a implementar no es clara, debido a que esto depende del tiempo y recursos que se dispongan, ya que el modelo predictivo más preciso es de alto costo en tiempo construcción, ejecución y requiere de varios pasos en comparación a la media móvil según lo evidenciado en esta investigación.

5.1 DISCUSIÓN

Debido a la diversa cantidad de decisiones que tomar y parámetros que decir a lo largo de este trabajo, se dejan un par de recomendaciones que pueden ser útiles al momento de realizar un experimento similar.

En esta investigación se decidió debido a criterios de negocio que el MAE sería la métrica que decidiría qué modelo era el mejor, pero este tiene ciertas desventajas como ser circunstancial, es decir, varía a través del tiempo, en este sentido sería interesante probar con métricas diferentes para decidir qué modelo es el mejor para cada serie, hacer un "Time series cross validation", considerando un promedio de la métrica a través del tiempo o directamente ver otra métrica como el RMSE.

En esta memoria solo se buscaba validar que el clasificador tuviera resultados aceptables, sin poner foco al clasificador utilizado, en futuros estudios se recomienda probar con diversos clasificadores y comparar con métricas de accuracy, f1-score, etc, para asegurar un buen clasificador, además hacer un mayor trabajo de feature engineering para sacar más provecho de las características obtenidas de las series de tiempo.

Dada la importancia del clasificador, se recomienda tener un mayor set de series de tiempo con historia suficiente, además de asegurar la limpieza de estas, para así confiar en una buena calidad de las series de tiempo, es decir, un buen set de entrenamiento para el clasificador, dado que, si el modelo es entrenado con data imprecisa o defectuosa, el resultado será deficiente (garbage in, garbage out), se estima que esta medida ayudaría a la precisión del clasificador, lo que de mejorar permitiría incorporar más etiquetas.

Se recomienda probar también el clasificador en series de tiempo no relacionadas con el rubro, es decir, ver el rendimiento del clasificador ya entrenado, al ingresar datos de competencias, de acciones, entre otros.

6 BIBLIOGRAFÍA

Bates, J., & Granger, C. (1969). The Combination of Forecasts: . *Operational Research Quarterly, 20(4),*, 451–468.

Brieman, L. (2001). Random forests. *Machine Learning 45(1)*, 5-32.

De Livera, A., Hyndman, R., & Snyder, R. (2011). Forecasting time series with complex seasonal patterns using exponential smoothing. *J American Statistical Association*, 11-16.

Fan, S., & Hyndman, R. J. (2012). Short-term load forecasting based on a semi-parametric additive model. *IEEE Transactions on Power Systems, 27*(1), 134-141.

Holt, C. C. (2004). Forecasting seasonals and trends by exponentially weighted. *international journal of forecasting 20*, 5-10.

Shu Fan, S. M. (2012). *robjhyndman .* Retrieved from https://robjhyndman.com

Talagala, T., Hyndman, R., & Athanasopoulos, G. (2018). Meta-learning how to forecast. *Department of Econometrics and Business Statistics*.