UNIVERSIDAD DE LOS ANDES

DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

MIIA-4203 MODELOS AVANZADOS PARA ANÁLISIS DE DATOS II

GRUPO 15: Luis Jorge Garcia Camargo **y** Fernando Pérez Moreno

Micro proyecto 1: *start up* agrícola

I. Introducción: definición del problema y la pregunta de investigación

El negocio de esta start-up está en ser el único intermediario entre los agricultores y consumidores finales de productos nativos del campo. La oferta de valor es llegar al cliente con un producto en el momento en que este lo requiere garantizando siempre la característica más importante: su frescura.

Los problemas que se presentan en el negocio están asociados principalmente a las pérdidas en dos vías: por un lado, los productos que no se venden a tiempo, al ser perecederos, se dañan; generando desperdicio alimentario y pérdidas en relación con los costos de adquisición, almacenamiento, transporte y de producción. Por otra parte, no contar con suficiente inventario de productos cuando el cliente demanda, genera insatisfacción y pérdida de ingresos potenciales.

La pregunta de negocio de la start-up está principalmente asociada a la política de la cadena de suministro: cuánto y cuándo pedir cada uno de los productos nativos a cada uno de los proveedores. Para responder a esta pregunta, es importante revisar el comportamiento de la demanda de los productos y buscar hacer una estimación de esta en la medida de los posible.

El presente proyecto buscar el desarrollo de modelos que permitan pronosticar la demanda semanal de los productos agrícolas de tal manera que apoye la toma de decisiones de la start-up frente al manejo de la cadena de suministro. La hipótesis del proyecto es que existe al menos un modelo que permita pronosticar la demanda para los productos más relevantes para al start-up en cuanto a su frecuencia de compra que cuenten con la suficiente información estadística histórica.

II. Metodología desarrollada

La metodología desarrollada en el proyecto llevó a cabo los siguientes pasos:

1. **Exploración de la información**: entendimiento de los datos históricos.
2. **Preparación de los datos**: agrupación de los datos, selección de los productos con suficiente información para el pronóstico, imputación de los datos faltantes después de la selección; la estrategia acá es pronosticar la demanda de los productos con mayor frecuencia de venta, mas no los productos con mayor volumen de venta.
3. **Calibración de los modelos de pronóstico**: se proponen cuatro modelos con una evaluación dinámica (*rolling forecast)* de una semana adelante para los últimos 8 periodos de cada serie, y para cada uno se calcula la desviación estándar de los residuales (RMSE) como criterio de selección y se escoge el modelo con el mejor desempeño para cada uno de los productos.
4. **Pronosticar de demanda**: para las próximas cuatro semanas para los productos relevantes.
5. **Concluir y recomendar** a la start-up con base en los hallazgos en del proyecto.

III. Resultados y análisis

A continuación, se resaltan los principales hallazgos en cada uno de los pasos de la metodología propuesta. Cabe a aclarar que toda lo que se indica a continuación tiene como base el notebook de Python anexo al presente proyecto.

1. **Exploración de la información**:

ID	Cliente	Fecha	Pedido	Precio	Producto	Nombre
0	Cliente26	18/09/2017	20	700	VER0049	Yerbabuena / 100 gramos

Tabla 1: registro 1 de la base de datos del cliente

La base contiene 3.971 registros con 6 variables: cliente, fecha, cantidad de pedido, precio y la identificación del producto con código y nombre (con su presentación). La información está por día desde noviembre de 2016 hasta septiembre de 2017, contando con 47 semanas seguidas. Además, existe información diversa para 121 productos nativos.

2. **Preparación de los datos**:

- **Definición de variables relevantes**: el proyecto se concentrará en el pronóstico de la demanda de los productos en cuanto a la cantidad de venta independientemente de su precio o del cliente que lo compre. Por lo tanto, solamente se trabajará con la fecha, nombre de producto y pedido.
- **Agrupación de los datos:** los datos se agrupan de manera semanal dado que el pronóstico se entregará de esta manera. Además, no todos los productos cuentan con registros diarios de pedido por lo tanto es muy poca la información para hacer pronósticos diarios. Así mismo, una agrupación mensual nos permitiría trabajar con más productos, pero tendríamos a lo más 11 puntos en cada serie, lo cual no sería conveniente usar ya que, según la teoría clásica de series de tiempo, un modelo debería estimarse con al menos 50 observaciones equidistantes en el tiempo.
- **Selección de los productos más relevantes para pronóstico:** se aplica un criterio de selección para determinar cuáles productos son susceptibles para la aplicación de los modelos de pronóstico con base a la cantidad de información nula. Entre mayor sea la cantidad de datos nulos se minimizan las posibilidades de desarrollar modelos de pronóstico. Por lo tanto, solo se seleccionan los productos que tienen hasta el 30% de las semanas (aproximadamente 14 de 47) de información faltante. Solo 18 de 121 productos cumplen con el criterio: Banano Criollo, Cebolla Cabezona, Roja Cebolla Larga, Champiñón, Cilantro, Espinaca, Fresa Pareja, Lechuga Crespa, Limón Tahití, Mango Tommy Atkins, Mora Castilla, Pepino Cohombro, Perejil Liso, Pimentón Rojo, Piña Golden Sweet, Tomate Chonto, Zanahoria y Zucchini Verde.
- **Imputación de los datos faltantes:** con la ayuda de la función *interpolate* haciendo uso del método *'time'* se hace uso de una estimación lineal para imputar la cantidad pedida en la semana que aparece con información nula.

Las series para los 18 productos más relevantes con base en lo desarrollado anteriormente se pueden ver en la Gráfica 1 del Anexo. A simple vista se pueden observar series con alto grado de aleatoriedad, con ciclos poco claros, y una aparente no correlación tan marcada entre los productos. Por tal razón, se decide que para cada producto se buscará el mejor modelo de pronóstico de manera independiente.

3. **Calibración de los modelos de pronóstico**: se seleccionan los siguientes cuatro modelos:

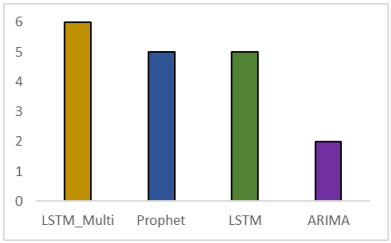
a. AutoARIMA: Es un modelo estadístico que utiliza variaciones y regresiones de datos estadísticos con el fin de encontrar patrones para una predicción hacia el futuro. Se trata de un modelo dinámico de series temporales, es decir, las estimaciones futuras vienen explicadas por los datos del pasado y no por variables independientes; este se vuelve a estimar conforme entran nuevas observaciones reales.
b. Facebook Prophet: Es una biblioteca de código abierto para el pronóstico de series temporales univariadas (una variable) desarrollada por Facebook. Prophet implementa lo que ellos llaman un modelo aditivo de predicción de series de tiempo, y la implementación apoya las tendencias, la estacionalidad y los días festivos.
c. Red neuronal LSTM univariado: La característica principal de las redes recurrentes es que la información puede persistir introduciendo bucles en el diagrama de la red, por lo que básicamente, pueden "recordar" estados previos y utilizar esta información para decidir cuál será el siguiente; utilizando solamente la estructura de correlación de cada serie
d. Red neuronal LSTM multivariado: Análogo al caso univariado, para este modelo multivariado se hará uso de estructuras de correlación más complejas ya que para el pronóstico de un producto *i* se tendrá en cuenta la información que aportan los demás productos de manera conjunta.

Con el fin de seleccionar el mejor modelo se calculó la desviación estándar de los residuales (RMSE) como criterio de selección. En el Anexo B se pueden observar las gráficas que comparan los pronósticos versus los datos reales del test. En resumen, los resultados son los siguientes:

		RMSE						Pronóstico (unidades)			
Producto	**Pres.**	**ARIMA**	**Prophet**	**LSTM**	**LSTM_ Multi**	**Modelo con RMSE más bajo**	**RMSE Mejor**	**1/10/17**	**8/10/17**	**15/10/17**	**22/10/17**
Cilantro	100 g	10.36	10.20	5.09	5.01	**LSTM_Multi**	**5.01**	7.63	8.29	8.30	8.29
Perejil Liso	100 g	8.16	12.31	8.64	7.14	**LSTM_Multi**	**7.14**	4.56	4.14	4.12	4.14
Limón Tahití	Libra	7.95	7.94	6.13	5.92	**LSTM_Multi**	**5.92**	14.84	16.14	16.17	16.15
Cebolla Cabezona Roja	Libra	19.10	18.17	17.35	16.51	**LSTM_Multi**	**16.51**	23.29	24.62	24.65	24.66
Tomate Chonto	Libra	38.87	39.26	37.33	37.14	**LSTM_Multi**	**37.14**	74.61	77.13	77.19	77.23
Piña Golden Sweet	Unidad	2.28	8.62	2.20	1.89	**LSTM_Multi**	**1.89**	4.25	4.52	4.53	4.53
Fresa Pareja	Libra	6.64	6.08	7.02	6.92	**Prophet**	**6.08**	5.49	5.52	5.54	5.56
Cebolla Larga	Libra	7.28	6.09	7.65	7.52	**Prophet**	**6.09**	4.02	3.99	3.97	3.94
Pepino Cohombro	Libra	11.38	8.28	8.68	11.46	**Prophet**	**8.28**	7.74	6.29	4.85	3.40
Champiñón	Libra	11.61	10.30	11.31	10.59	**Prophet**	**10.30**	17.68	18.06	18.45	18.83
Lechuga Crespa	Unidad	17.17	13.95	14.38	14.31	**Prophet**	**13.95**	36.59	37.14	37.70	38.25
Mora Castilla	Libra	3.44	10.23	3.08	3.22	**LSTM**	**3.08**	3.47	3.67	3.80	3.92
Mango Tommy Atkins	Libra	5.04	7.90	3.72	3.79	**LSTM**	**3.72**	7.04	6.88	6.75	6.66
Banano Criollo	Libra	15.09	11.30	10.28	11.84	**LSTM**	**10.28**	21.44	21.30	21.19	21.11
Zucchini Verde	Libra	20.89	26.05	19.20	20.50	**LSTM**	**19.20**	27.36	25.22	23.32	21.99
Zanahoria	Libra	50.23	61.14	48.08	62.41	**LSTM**	**48.08**	86.30	92.50	95.71	99.08
Espinaca	Libra	2.54	8.78	3.30	3.05	**ARIMA**	**2.54**	1.80	2.65	2.92	4.01
Pimentón Rojo	Libra	33.29	54.10	43.40	53.09	**ARIMA**	**33.29**	48.02	21.47	25.27	27.61

Tabla 2: Resultados de los modelos y pronóstico de 4 semanas

La Tabla 2 muestra los 18 productos más relevantes, el resultado del RMSE para cada modelo, el mejor para cada producto y se presentan los pronósticos para las siguientes 4 semanas que se analizará en la siguiente sección. En general se puede ver que no siempre el mejor modelo es el mismo para cada producto. Por el contrario, para los 18 productos la frecuencia del mejor modelo fue bastante dispersa:



Grafica 1. Distribución de las veces en las que resultó mejor modelo

El mejor modelo es en seis ocasiones la red neuronal LSTM multivariada, se igualan en frecuencia de cinco veces prophet y la red LSTM univariada, y el auto ARIMA resulta el mejor en dos casos. Lo anterior muestra que siempre se debe probar con varios modelos y definir un criterio de comparación universal para escoger el mejor.

4. **Pronosticar de demanda**:

Como parte de la consultoría se le entrega a la start-up agrícola el mejor pronóstico para las próximas 4 semanas con base en el modelo que presentó el mejor desempeño relativo entre los 4 modelos para cada producto. Los pronósticos se entregan con dos decimales, y se pueden observar en la Tabla 2. Por otra parte, en el Anexo C se muestran las gráficas de cada pronóstico.

IV. Conclusiones y recomendaciones

Si bien la start-up ha generado ventas en 121 productos diferentes no para todos es posible construir un modelo de pronóstico. Esto se debe a que es necesaria una mínima cantidad de datos que permita entrenar diversos modelos y posteriormente poder probarlos y compararlos para escoger el mejor. Así las cosas, se agregaron los datos diarios en semanas y se encontró información suficiente para desarrollar modelos de pronóstico para 18 productos.

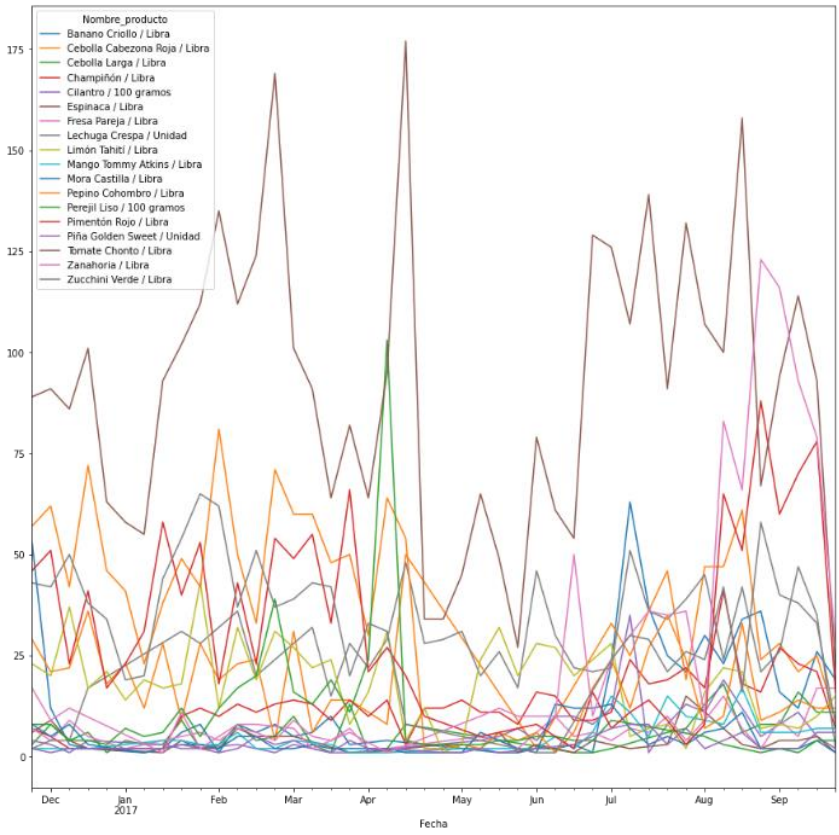
Se encontró que el comportamiento de los 18 productos es muy diferente, y al probar 4 modelos diferentes en cada uno, se pudo escoger el mejor con base en el criterio de la menor desviación estándar de los residuales RMSE. Resultó que para 6 productos el mejor modelo es la red neuronal LSTM multivariada, para 5 el modelo de Facebook prophet, para 5 la LSTM sencilla mientras que para 2 el ARIMA. Lo que nos permite concluir que el mejor modelo relativo depende del tipo de producto.

Es importante que la start-up establezca un modelo de recolección de datos más robusto. Con la información actual no se pueden observar estacionalidades ni tendencias claras, la series con información semanal no permiten observar un comportamiento donde sea más fácil de entender el ciclo. Este esfuerzo debe focalizarse en los productos con mayor frecuencia de demanda que obliguen a tener stocks permanentes. No es necesario desgastar esfuerzos en productos que se demandan de manera puntual o esporádica.

También se sugiere incluir otras variables externas que permitan planear la cadena de suministro. Por ejemplo, la información de la época del tiempo, estaciones y ciclo de productos y cosechas. También se puede incluir información secundaria de cuándo los productos tienen mayor demanda en general. Así, se podrían establecer demandas mensuales siempre y cuando la información recolectada sea lo suficientemente robusta.

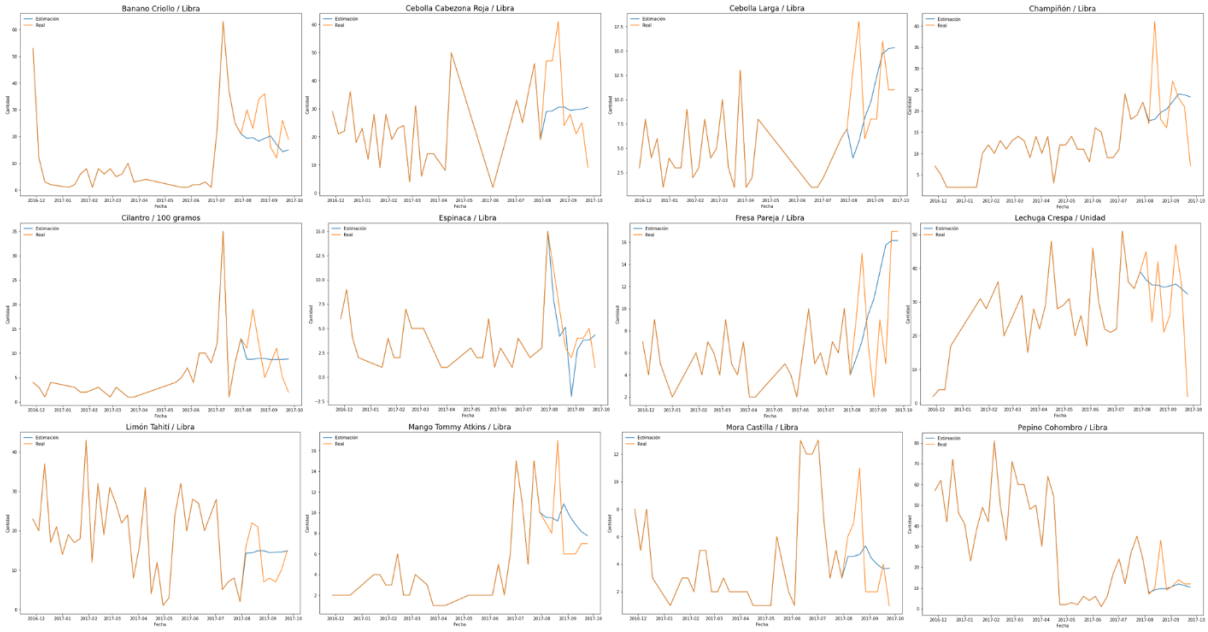
V. Anexos

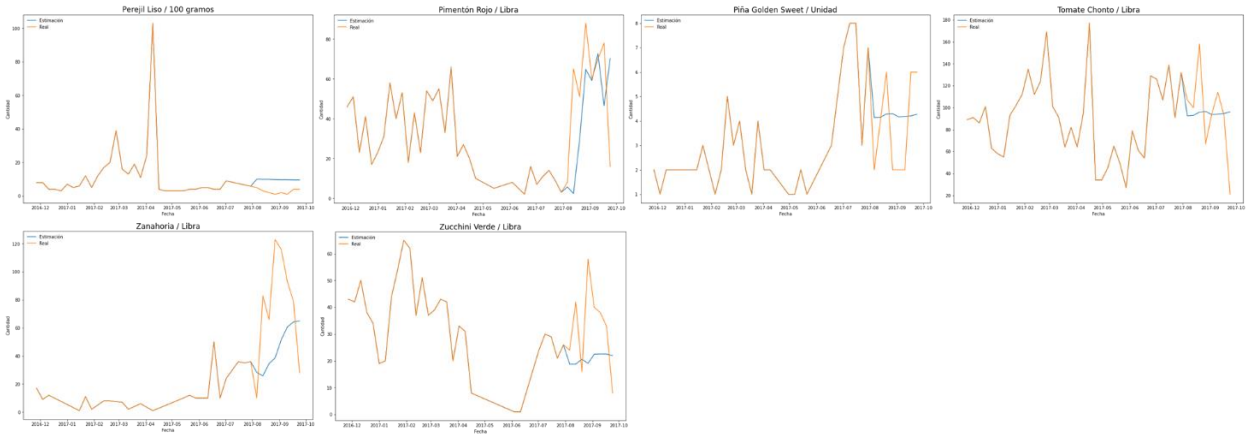
Anexo A. Serie semanal de los 18 productos más relevantes:



Gráfica 1: Serie de pedidos para 44 semanas de 18 productos relevantes

Anexo B. Los pronósticos del test se pueden observar para cada uno de los 18 productos:





Anexo C. Los pronósticos para 4 semanas de los 18 productos más relevantes: