Imagen que contiene dibujo, plato

Descripción generada automáticamente DESAFÍO DATA SCIENTIST

Nombre Candidato: Binicio Andrés Barrueco Uauy

Contacto: biniciobarrueco@protonmail.com

Una cadena de retail quiere predecir sus ventas a lo largo del tiempo. Para ello, han juntado sus datos históricos y nos han contratado para armar un modelo de forecast.

Los datos incluyen información como el id de la boleta, la fecha, id del producto, descripción del producto, precio, entre otras. Lamentablemente, no cuentan con un diccionario de variables, por lo que es tarea del Data Scientist entender las variables y dar una pequeña descripción en el desarrollo del ejercicio.

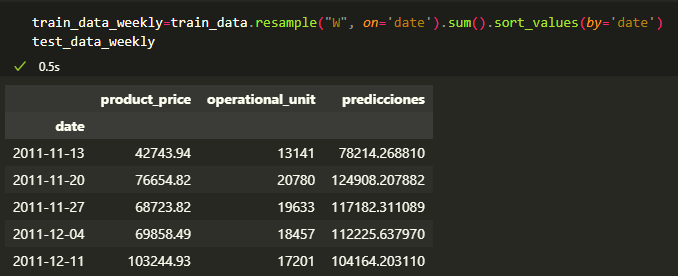
El objetivo del ejercicio es:

1. Entender los datos y construir el mejor modelo que sea capaz de predecir las cantidades vendidas a futuro. El modelo queda a criterio del Data Scientist. La predicción será semanal, y se piden 4 semanas de predicción.
2. Se deja a criterio del data scientist la métrica, pensando en lo que busca el cliente.
3. Entregar las respuestas del dataset test.csv como archivo csv (como columna extra del test.csv).
4. ¿De qué manera se puede complementar la solución? Pensar en propuestas para el cliente.
5. ¿Cuáles pueden ser las limitaciones, riesgos, sesgos de los modelos al implementar este tipo de soluciones? ¿Qué otras cosas hay que considerar al momento de implementar un proyecto como este?
6. Los resultados deben ser entregados en un Jupyter Notebook autocontenido.

Tendrás un plazo de 1 semana para entregar el caso.

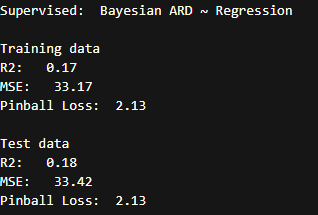
***Desarrollo y resultados en las hojas siguientes***…

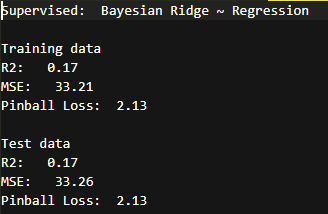
(P1) Modelo Construido y utilizado: “*Linear Regression”*

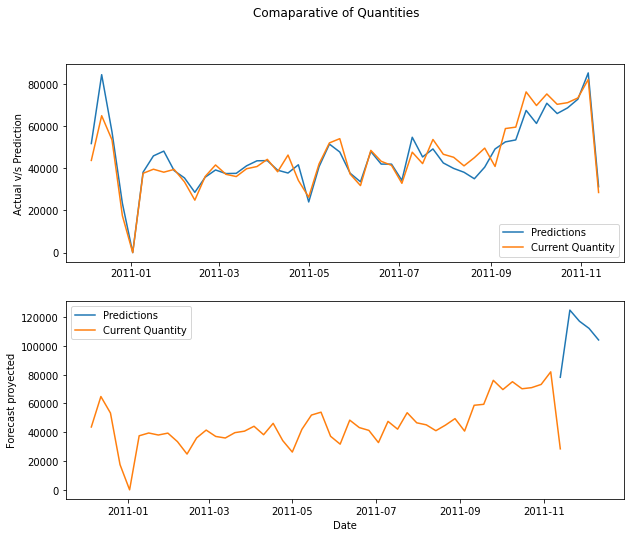
***Predicción semanal para los siguientes períodos*:**

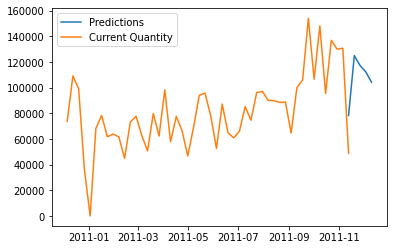
Para desarrollar el modelo se utilizó una regresión lineal, la cual fue comparada con otros modelos. Por criterio métrico y para efectos de acortar el desarrollo, nos basamos principalmente en el Error Cuadrado Medio (MSE) arrojado en el “test”, el cual busca minimizar el error entre las distancias entre la línea ajustada. Si bien todos los modelos, entregan buen rendimiento. La variabilidad es muy mínima. Para discriminar de mejor forma habría que agregar hiperparámetros y reentrenar los modelos.

***Resultado MSE= 33.16***

Texto

Descripción generada automáticamente

 (P4) Complemento de solución: Gráfico visual



El primer gráfico nos muestra lo actual versus lo predicho y el margen de error que hubo a simple vista.

Si observamos bien la escala de los datos, el gráfico Actual v/s Prediction está más filtrado que los otros, esto debido a la restricción fijada al comienzo con el rango intercuartílico y su límite superior. Al recortar los datos descartamos una cantidad considerable. En los dos gráficos inferiores se observa la comparación de escala con respecto a la predicción. Podríamos expresar esto como un sesgo y un posible error, para esto habría que explorar de forma exhaustiva los datos y ver si realmente son casos puntuales que vale la pena descartar.

En resumen, para complementar la solución, se observa que a fin del año 2010 la tendencia se proyecta en aumento, pero a comienzos de año 2011 desciende, el aumento a fin de año puede deberse al período navideño.

A Principios de año hay tendencia bajista, quizá debido a que los clientes ahorran para vacaciones, pagos de tasas de impuesto, entre otros gastos comúnmente realizados en estos períodos, lo que implica menores ventas.

A finales de enero del 2011 la curva nuevamente se inclina de forma alcista hasta unos meses antes del período proyectado. Como conclusión claramente se ve estacionalidad en algunos períodos.

* Peaks más relevantes:

Comienzos de febrero; septiembre a octubre y noviembre a diciembre

En función del tipo de retail que sea, en períodos navideños implementar estrategias para tener productos complementarios a los productos estacionales más vendidos en esas semanas. e.g: Visualmente en diciembre hay muchos juguetes vendidos; podrían crear promociones que añadan “pets supplies”, ya que, muchos clientes han adquirido hábitos de hacer regalos a sus mascotas.

* Valles más relevantes:

Comienzos de enero, mayo y julio. Y, fines de octubre.

Por lo tanto, sería bueno saber a que país pertenece el retail, varía su estación del año, período escolar, entre otros. En base al período escolar, e.g pueden existir pack de productos que por una suma adicional incluyan tablets. Además, como supuesto están los seguros obligatorios a comienzos de año. Sería bueno crear alianzas para entregar descuentos o bien gratuidad de este seguro por presentar boleta con compras superiores a cierto monto.

(P5)

* En base a las columnas que tenemos disponibles, puede que no sean suficientes características para que los modelos comparen y correlacionen e impliquen causalidad una en la otra. Quizá sería bueno tener clasificados a los tipos de clientes para incluir mejores predicciones. e.g. recurrente, y no frecuente.
* Los datos seleccionados para el pronóstico no sean lo suficientemente aleatorios para lograr equidad en las muestras, vale decir; dado que descartamos cantidades que estaban por sobre 26, puede que en la BBDD de donde provienen las muestras haya más como las descartadas, e.g sería como descartar personas con estaturas mayores a 1.7 y dejamos solo las de 1.69 hacia abajo.

Estaríamos sesgando la selección y dejando fuera más observaciones.

* Al analizar múltiples datos no conocidos, desde el punto de vista del negocio, el analista podría realizar un análisis exploratorio con mala interpretación.
* En cuanto a los riesgos, si un modelo es creado bajo un escenario de variables estático, esto puede generar errores en un futuro: Si la empresa o bien el cliente que solicita el modelo cambia sus estrategias o busca crecer, e.g apertura nuevos negocios que pueden afectar sus resultados de las ventas, los modelos no estarían preparados para ajustarse a esos cambios.