Resumen Ejecutivo

Proyecto: Predict Future Sales
"How to win a data science competition" Coursera course

Equipo:

- Chief Exploratory Data Scientist: **Alejandra Lelo de Larrea Ibarra** 124433
- Chief Machine Learning Scientist: **Dante Ruiz Martínez** 183340
- Chief Data Engineer and Project Manager: **José Carlos Escobar Gutiérrez** 175895

Problemática:

Pronosticar las ventas mensuales por producto y tienda para la compañía de software 1C de las más grandes de Rusia. El propósito de estas predicciones es informar la implementación de estrategias de negocio, campañas de marketing, correcta administración de inventarios e identificación de áreas de oportunidad.

Objetivo General:

Desarrollar un producto de ciencia de datos para predecir el nivel de ventas mensuales por tienda y artículo para la de la empresa 1C utilizando la metodología CRISP-DM. El criterio de éxito de este producto es superar el benchmark de 1.67 en la métrica, raíz de la suma de los residuos al cuadrado (RMSE) de la predicción de ventas por producto y tienda en el mes de noviembre de 2015. El periodo de análisis abarca de enero de 2013 a octubre de 2015. Los datos fueron proporcionados en la competencia de Kaggle: “Predict Future Sales. How to Win a Data Science Competition”.

Objetivos Particulares:

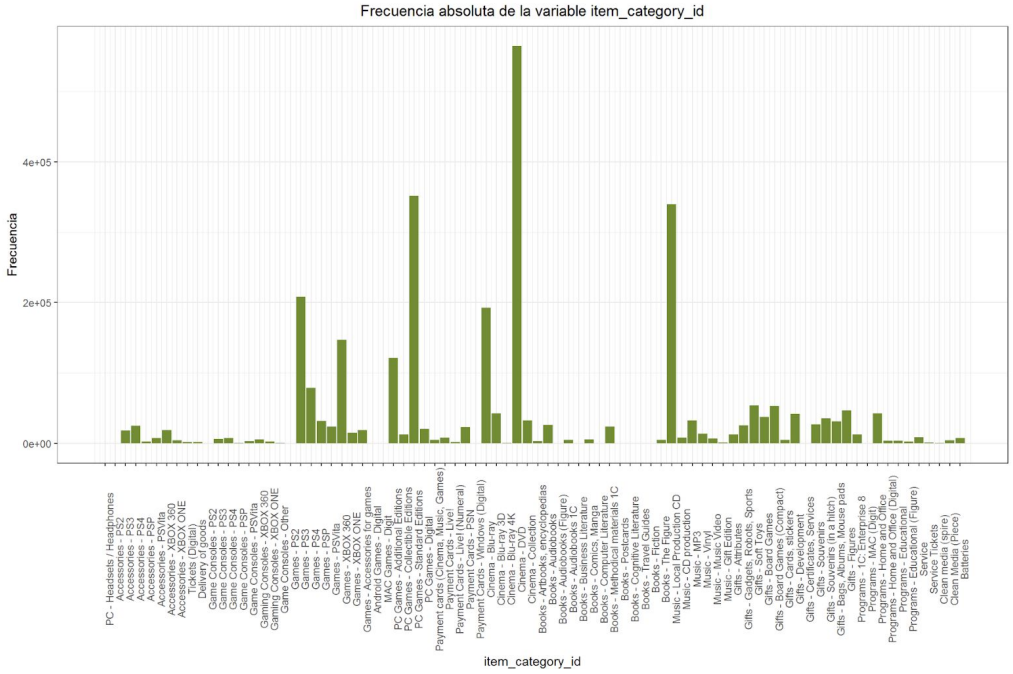
- **Entender el negocio.** Desarrollar herramientas para adquirir y analizar los datos sobre las ventas de la compañía.
- **Preparación de los datos.** Construir una base de datos limpia y con ingeniería de características que sirva para modelar el comportamiento de las ventas por producto y tienda.
- **Modelado**. Seleccionar y ajustar varios algoritmos sobre un conjunto de entrenamiento y predecir en un conjunto de prueba.
- **Evaluación**. Comparar el desempeño de varios modelos y escoger el mejor para predecir el nivel de ventas por producto y tienda.
- **Implementación del proyecto**. Desarrollar un app para generar predicciones utilizando el mejor modelo

Principales Resultados:

a) Principales resultados del análisis exploratorio de datos.

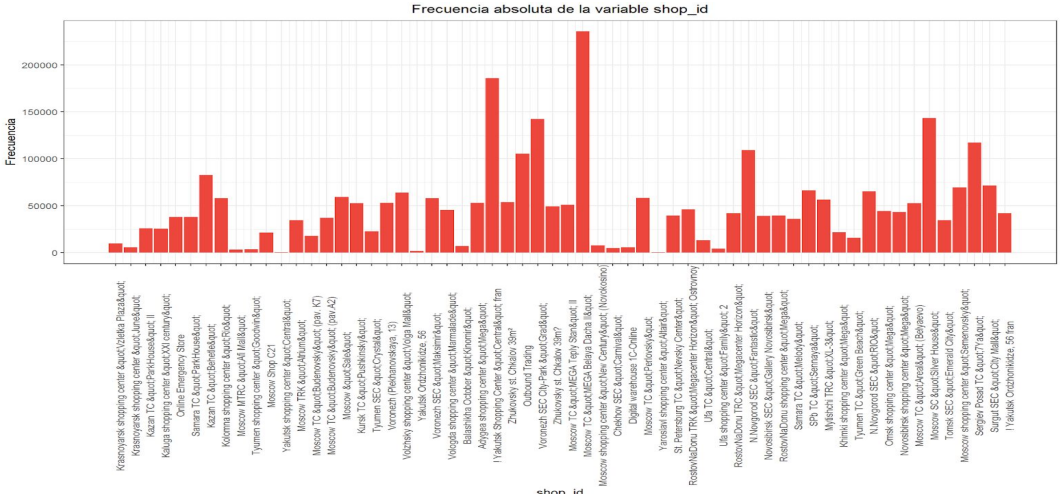
- Los artículos más vendidos son aquellos que pertenecen a la rama de entretenimiento: películas, videojuegos, consolas, música. Ver figura 1.
- Los artículos menos vendidos son los que pertenecen a la rama de educación. Ver Figura 1.

Figura 1. Los productos más vendidos son películas, videojuegos, consolas y música.



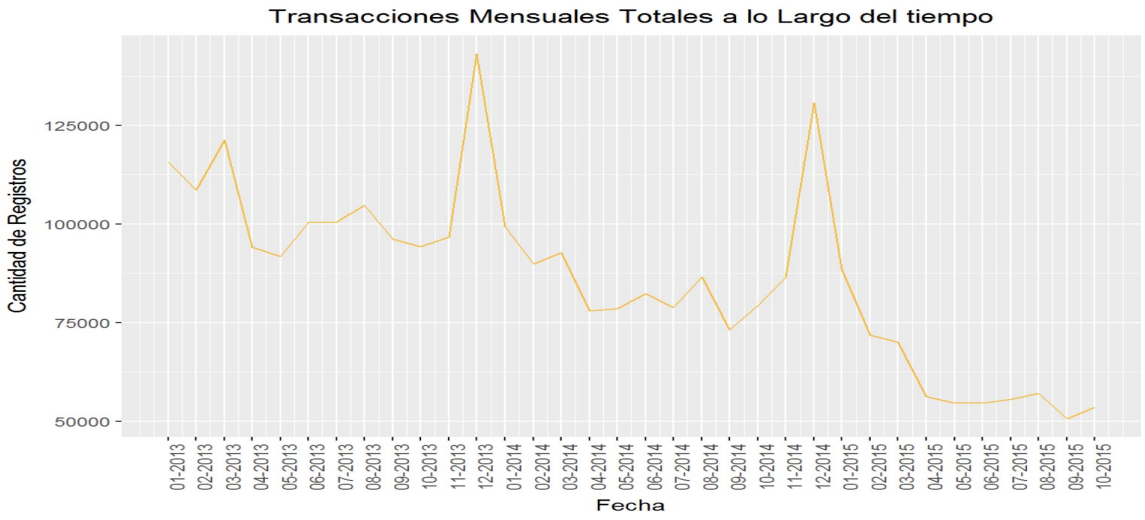
- Presumiblemente, las tiendas con mayor número de transacciones son las que están en Moscú. Ver figura 2.

Figura 2. Las tiendas con mayores ventas están en Moscú



- Las ventas de 1C Company han disminuido a lo largo de la muestra. Ver figura 3.
- Se tienen picos (ventas extraordinarias) en los meses de diciembre. Ver figura 3.

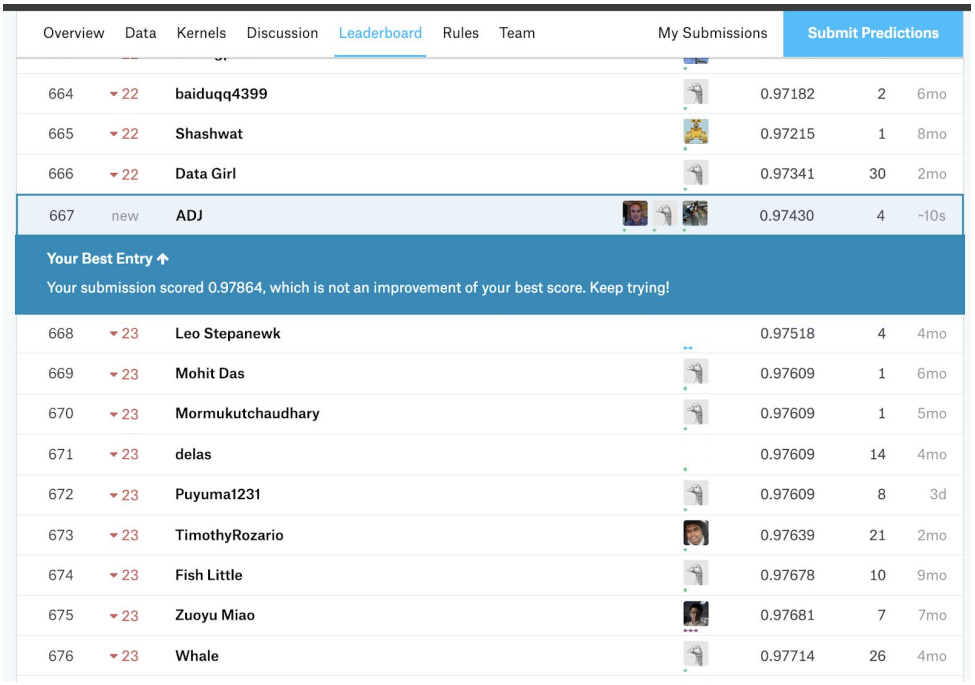
Figura 3. Las ventas tienen una tendencia negativa y picos en los meses de diciembre



b) Modelado de predicción de ventas por producto y tienda de manera mensual

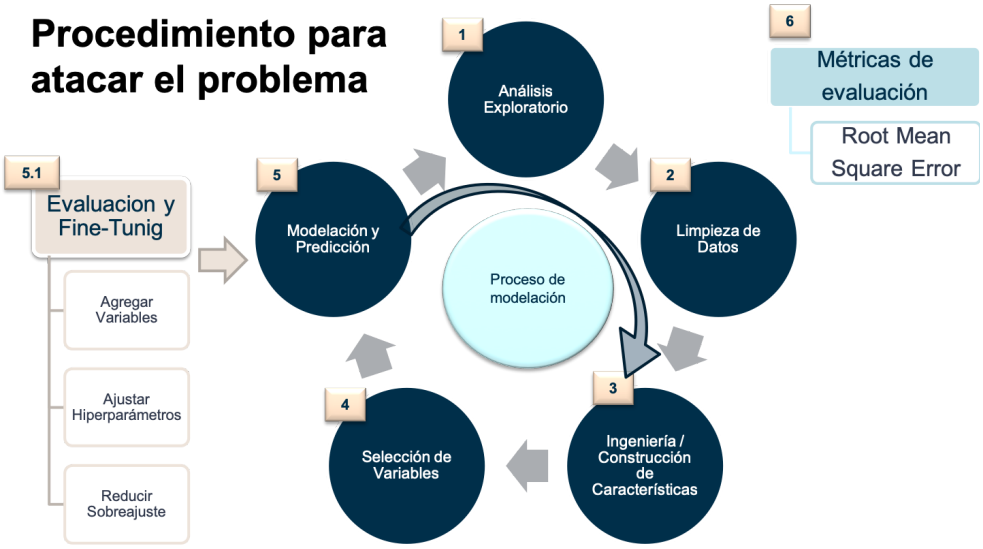
- Se hicieron tres modelos ridge, lasso y xgboost. El que mejor se desempeñó fue el XGBOOST el cual superó en todas las ocasiones el benchmark de 1.67 en el RMSE sobre el conjunto de prueba. La predicción más alta que se obtuvo fue de 0.97 RMSE colocando el modelo en el lugar 667 de 1986. Ver figura 4.

Figura 4. El mejor modelo superó el criterio de éxito del proyecto alcanzando un 0.97 del RMSE



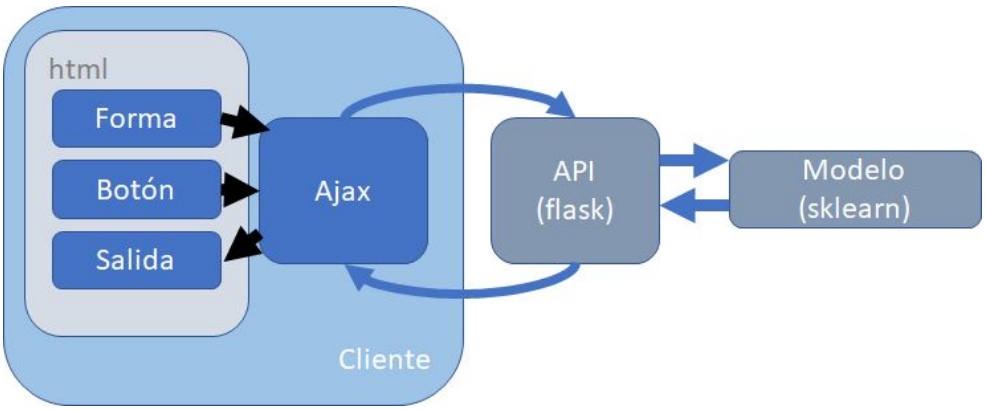
- El procedimiento que se utilizó para obtener los modelos se basó en la metodología CRISP-DM. Una vez entendidas las necesidades del negocio de procedió a armar una base de datos, realizar un análisis exploratorio de datos, se generó una base de datos limpia y con nuevas características, se seleccionaron las variables necesarias para predecir en el futuro y se ajustaron modelos. Los mejores modelos fueron probados en Kaggle para verificar su desempeño y proceder a su optimización. Ver Figura 5.

Figura 5. Metodología utilizada en el desarrollo del proyecto



c) Herramienta de flask para predecir ventas.

Flask es un microframework que permite crear aplicaciones web con pocas líneas de código, se enfoca en proporcionar lo mínimo necesario para poner a funcionar una aplicación básica, por ejemplo, para el prototipado rápido de proyectos. Incluye un servidor web de desarrollo para prueba de aplicaciones sin tener que instalar mas aplicaciones. Flask cuenta con un depurador y soporte integrado para pruebas unitarias así como soporte para Unicode y es compatible con WSGI 1.0.Esta manera de operar es ideal para poder implementar aplicaciones en un celular o bien a través de una página web, por ejemplo.



Retos enfrentados y a considerar a futuro:

El principal reto de este problema fue la limpieza de los datos ya que estos se encontraban en idioma Ruso. Por lo que se utilizó la API de Google Translate para traducir la información que fue clave para el proceso de entender el negocio, por ejemplo, qué productos, categoría de productos y tiendas había en la base de datos.

Fue necesario completar la base de datos para que el algoritmo pudiera aprender el patrón del tiempo. Las series se completaron construyendo todas las combinaciones de tienda y artículo por mes. Asimismo, se quitaron los precios negativos que correspondían a las devoluciones y se corrigieron los precios que no traían punto decimal.

Otro reto es el de generar una base de datos con todas las características nuevas para predecir hacia adelante. Toda esta información se genera a partir de la variable de ventas mensuales por tienda y artículo.

Consideraciones finales:

El producto de ciencia de datos cumplió satisfactoriamente con los requisitos del proyecto pues superó la métrica de éxito establecida por la compañía 1C. El modelo desarrollado en este proyecto para la compañía superó el benchmark de 1.167 en el RMSE. Dicho modelo se implementó en una aplicación de Flask para predecir el nivel de ventas por producto y tienda. El resultado de este proyecto es reproducible, actualizable y con posibilidades de optimizarse más.

Implicaciones para el negocio:

El producto de ciencia de datos que se desarrolló en este proyecto está apto y listo para implementarse en la compañía 1C. La herramienta predictiva es capaz de proporcionar mejores predicciones sobre el nivel de ventas mensuales de productos por cada tienda. Esto sería de gran utilidad para implementar una estrategia integral para elevar las ventas de la compañía que durante el periodo de estudio se observó venían cayendo. En ese sentido, el modelo puede servir de insumo para generar una campaña de marketing focalizada en tiendas y productos particulares. Asimismo, puede servir para optimizar los canales de distribución y mantener en niveles adecuados los inventarios de mercancía. Por otro lado, también puede ser utilizado para estudiar mejor el comportamiento agregado de los cliente ante diferentes eventualidades como periodos vacacionales, días festivos, periodos de ofertas, etc.

Lista de Entregables:

La siguiente lista contiene un inventario de entregables que se ponen a disposición en el repositorio de github de este proyecto: https://github.com/texantubber/Final_Mineria. Los archivos se enlistan por orden de relevancia.

Presentables:

1. Comprension_Datos.html: Página web que describe los datos utilizados
2. Traduccion.html : Página web que describe el proceso para traducir la base de datos
3. Entendiendo_el_Negocio : Página web que describe el problema del negocio
4. EDA.html: Página web que presenta el análisis exploratorio de datos
5. Preparacion_Datos.html : Página web que describe cómo se llevó a cabo la limpieza de los datos y la ingeniería de características.

6. Modelado.ipynb : Página que describe cómo se llevó a cabo el proceso de modelado.
7. Evaluacion.ipynb : Página que describe cómo se evaluaron los modelos en Kaggle

Código reproducible:

8. README.md: Contiene instrucciones de cómo hacer reproducible el producto de ciencia de datos.
9. DescargaDatos.sh : Código bash para descargar archivos
10. Comprension_Datos.Rmd
11. Traduccion.Rmd
12. Entendiendo_el_Negocio.Rmd
13. EDA.Rmd
14. Preparacion_Datos.Rmd
15. 01_Libraries.R: Script de R con toda la paquetería necesaria para reproducir el producto.
16. 02_Utils.R: Script de R con herramientas para procesar los datos.

App

17. predictor.py: el código que se monta como app y funciona como servicio web para predecir ventas.

Carpetas:

1. Datos: Contiene los archivos con los datos crudos.
2. Datos_trad: Aloja los datos traducidos del ruso al inglés
3. Datos_clean: Aloja los datos limpios
4. Data_Modelos: Aloja los datos del conjunto de entrenamiento y prueba de los modelos
5. Modelo_Final: Aloja los datos del modelo final
6. Flask: Aloja el código de la aplicación.