Carrera DataScience

Proyecto Final

Hito 1: Proyecto Automotora Anaconda

Integrantes

Jorge Guerrero
Daniel Mardones
Gonzalo Rojas
Esteban Sánchez
Paulo González

Profesores

Daniel Befferman
Jhon Poma

29 de Enero del 2023, Chile

Equipo de trabajo

Para llevar a cabo el presente proyecto, los integrantes del grupo han pre definido roles los cuales se enfocarán en áreas específicas del trabajao a modo de tener un orden al momento de trabajar. Los roles fueron definidos en base a las características de los integrantes y se presentan a continuación:

- Líderes: Jorge Guerrero – Paulo González
  - Encargados de mantener actualizados al equipo y plantear objetivos para finalizer las entregas.
- Analistas de datos: Esteban Sanchez – Gonzalo Rojas
  - Encargados de estudiar los datos disponibles y de enriquecer la base de datos para construir un modelo más robusto.
- Control de calidad y validación de datos: Esteban Sanchez – Gonzalo Rojas
  - Encargados de que los datos existentes o agregados mantengan el mismo orden, en términos de limpieza y codificación.
- Ingenieros de modelamiento: Daniel Mardones – Jorge Guerrero
  - Encargados de proponer, modelar y testear modelos para avanzar hacia algoritmo final.
- Visualización de datos: Daniel Mardones – Jorge Guerrero
  - Encargados de exhibir la información recopilada de manera sencilla y entendible para lectores e integrantes.
- Documentadores: Daniel Mardones – Paulo Gonzalez
  - Encargados de mantener registros de los cambios que se han realizado en los archivos, en conjunto con copias de versiones pasadas de los mismo.

Selección del tema y motivación

El proyecto surge de la necesidad de encontrar un modelo que sea capaz de conseguir los mejores negocios para las automotoras, donde la idea es que el modelo sea un apoyo para la toma de decisiones ayuda'ándoles a tener mejores rentabilidades. Para ello, el modelo debe ser capaz de identificar autos que estén bajo el precio de mercado para comprarlos a particulares y luego revenderlos en el mercado a un precio más elevado y permitiendo a la automotora ganar el márgen de diferencia entre precio de compra y venta. El uso sistemático de la herramienta que se pretende crear, permitirá que la o las empresas puedan identificar con mayor facilidad los negocios que si serán rentables antes de realizar la inversion, de esta forma se reduce la incertidumbre en la toma de decisiones y se mejora la rentabilidad por cada negocio que se realice. En síntesis, se busca hacer un modelo que pueda ayudar a las automotoras filtrando negocios que pueden ser posiblemente rentables de aquellos que no para luego armar estrategias de venta.

Requerimientos

Es importante tener en cuenta que para cumplir con las expectativas de este trabajo, se cuentan con los siguientes requerimientos para poder elaborar un proyecto completo. Los requerimientos se enumeran a continuación:

1. Generar una herramienta que ayude al equipo de compras a encontrar las mejores oportunidades, a través de un modelo de regresión que permita encontrar los autos a precios más baratos para luego venderlos en el mercado usando técnicas de Machine Learning. Lo que se busca con esta herramienta es que el vendedor tenga un primer filtro eficiente sobre cuales vehículos podrían generar un buen negocio para la automotra.
2. Definir una propuesta de inversión (lista de automóviles convenientes) para la Automotora. Por medio de las metricas obtenidos se podrá definir una propuesta para inversion, mediante vehiculos que esten por debajo del precio de venta promedio, lo cual permite un mejor margen de venta.
3. Definir una estrategia de venta por gama de vehículos. Se incorpora una columna la cual clasifica la marca del vehículo los cuales serán distinguidos como generalistas y premium. Esta decisión se basa en articulos que mencionan y hacen recuento de las marcas que se consideran como premium y aquellas que se consideran generalistas[1] .Con esto se puede generar un analisis del comporatamiento en la venta de vehiculos y con esto desarrollar una estrategia.
4. Encontrar oportunidades de compra y venta entre ciudades/estados. A través de un EDA, de podrá determinar en que ciudades / estados se encuentran las mejores oportunidades de compra y determinar las mejores ciudades / estados para vender estos vehiculos, maximizando la utilidad.

Planificación de la Investigación

Hipotesis

La hipotesis del proyecto se basa en rechazar que el modelo final que se planteé no consiga crear rentabilidades para la empresa. Rechazando la hipótesis nula anterior a través de los resultados que se obtengan del modelo se podrá verificar que el modelo es factible y rentable.

[1] Se puede encontrar los artículos en los siguientes enlaces: https://www.autopista.es/noticias-motor/las-marcas-de-coches-premium-y-generalistas-que-mejor-tratan-a-los-clientes_144781_102.html; https://www.expansion.com/fueradeserie/motor/2022/06/02/628f8853468aeb46118b4660.html

Definición del Vector Objetivo

El vector objetivo, el cual se buscará predecir corresponde a la variable contenida en la base de datos “Price”. Esta variable representa el valor de venta de los vehículos en dólares que se encuentran en una cierta ciudad, con cierto kilometraje y otros atributos que lo caracterizan.

Estrategias analíticas a nivel descriptivo

Se desarrollorá un EDA, el cual permitirá visualizer la distrubición de vehiculos de acuerdo a la marca, modelos, año, ciudad, estado, millas y valores de vehiculos. De esta forma, se puede entregar información más contundente sobre el comportamiento de los precios de los autos en cada uno de los atributos, para así tener un mayor conocimiento de las características más generales que se pueden extraer en la base de datos. Este análisis generará las líneas generales bajo los cuales se pueden elaborar, permitiendo observar que varialbes influyen más en lo niveles de precio de los autos y poder determinar estrategias de reventa o arbitraje entre ciudades una vez obtenido el precio estimado para un auto con ciertas características.

Modelación y predicción de trabajo

Como estrategia inicial para desarrollar la modelación del trabajo, se proponen tres modelos para probar resultados, estos son:

- Regresión Linear
- Decision Tree Regressor
- Linear GAM

Estos modelos se testearán sin hiperparámetros para luego agregar parámetros y luego ir observando sus resultados y de ser necesario, incluir más modelos al studio. Luego de la fase de prueba, se profundizará más en los modelos que tengan mejor desempeño, medido a través de las métricas de R-cuadrado y Error Cuadrático Medio. Es importante destacar los métricas utilizadas para medir el desempeño de los modelos del proyecto también corresponderán a R-cuadrado y el Error Cuadrático Medio.

Para almacenar los resultados que se obtienen de los modelos se hará uso de una base de datos Postgres, permitiendo así obtener Queries para verificar si la compra de un vehículo específico es conveniente.

Consideraciones

Es importante tener en cuenta algunos elementos que dictarán la manipulación de los datos obtenidos.

- En primer lugar, los datos no seran encriptados ya la información obtenida no contempla elementos que comprometan a terceros por lo que no serán censurados y serán usados libremente.
- En segudo lugar, es imprescindible mencionar que los datos con los cuales se trabajará corresponden a información entregada por la academia, la cual no presenta datos nulos y la cual se buscará enriquecer a lo largo del proyecto.
- Finalmente, existen 3 datasets los cuales por origen se encuentran divididos en:
    1. Muestra total = true_car_listings.csv
    2. Muestra entrenamiento = true_cars_train.csv
    3. Muestra validación = true_cars_test.csv

Debido a esto, no se realiza divisiones de muestra, solo de vector objetivo y matriz de atributos.