Ciencia de Datos Aplicada – Tercera entrega del proyecto

Jose F. Corzo Manrique, Felipe A. Gutiérrez Naranjo, Alejandro Mantilla Redondo
Reporte técnico del tercer *sprint* del proyecto
Universidad de los Andes, Bogotá, Colombia
{j.corzom, fa.gutierrez, a.mantillar}@uniandes.edu.co
Fecha de presentación: diciembre 11 de 2021

Tabla de Contenido

1 Enlaces relevantes de solución

Esta sección contiene los múltiples enlaces necesarios para soportar la segunda entrega del proyecto del curso:

- Repositorio específico donde se encuentra la solución de esta tarea https://github.com/MantiMantilla/Ciencia-de-datos-aplicada-2021-20/tree/main/Proyecto/Sprint%203

2 Entendimiento del negocio y enfoque analítico

El contexto del proyecto apunta a una clara necesidad de aumentar la adquisición de clientes para los taxistas en la ciudad de Nueva York, quienes son los clientes finales de esta solución. Se han detectado numerosos obstáculos en su quehacer laboral:

- Inseguridad vista en el nivel de violencia en su contra, al punto en que el conducir taxis es una de las profesiones más riesgosas de Nueva York.
- Usuarios que huyen sin pagar, representando trabajo hecho a costo personal.
- Competencia intensa por parte de aplicaciones como Uber o Lyft.
- El efecto de la pandemia de 2020.

Así, de define al problema como: **generar un plan de identificación de zonas y rutas que maximicen las ganancias de los taxistas, a partir del histórico de trayectos de taxi amarillo en la ciudad de Nueva York del TLC.**

Con base en lo anterior, la tarea predictiva específica en esta entrega es: **identificar las mejores métricas con las cuales predecir el mayor beneficio para los taxistas**. Como se explicará en las siguientes secciones, los mejores predictores identificados son los calificadores de días festivos y eventos de tormenta (obtenidos por enriquecimiento) junto a la hora del día y la zona en que se encuentran. A continuación, se ve un *Story board* actualizado donde se ve un ejemplo de la situación esperada de desarrollar la herramienta en su completitud, con un taxista de Nueva York ingresando los datos relevantes al *dashboard*, el cual calcula la zona con la mayor probabilidad de garantizarle una carrera para ese momento, y efectivamente consigue que alguien solicite su servicio.

El contar con esta herramienta en la vida real sería un haz bajo la manga para alcanzar un nuevo nivel competitivo con los demás prestadores de servicio de conducción, junto a la hipótesis que el identificar este tipo de rutas óptimas descartaría por defecto aquellos lugares donde se encuentra violencia y/o robo, puesto que serían económicamente despreciables. Ha de notarse que el despliegue visto en esta entrega es un prototipo del modelo final, en un archivo creado en *Jupyter Notebook*, del modelo en que se basaría el producto de llevar el proyecto más allá de la clase.

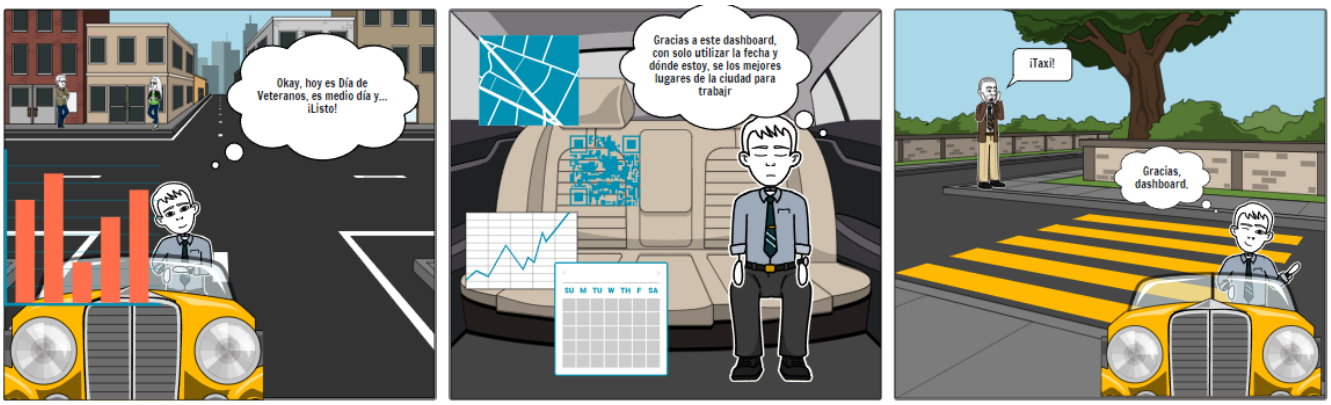

Figura 1. Storyboard de aplicación de solución

3 Entendimiento y preparación de los datos

Se realizan varios cambios frente a la entrega del segundo sprint, los cuales se detallan en las siguientes subsecciones. La variable objetivo del ejercicio continúa siendo “Total_amount” (representativo de la sumatoria del costo final que el (los) pasajero(s) paga al concluir el trayecto), mientras que las demás variables que forman parte del modelo son “total_time_min”, “total_trip_distance”, “total_fare_amount”, “total_tip”, “total_toll”, “total_amount”, “total_taxis” (todas el resultado de agregaciones de las variables descritas en el documento anterior). De modo similar, las variables de entrada son “year”, “month”, “dayofweek”, “dayofmonth”, “hour” y “PULocationID”.

3.1 Alteraciones en las variables para mejorar *dataset* original

Se identificaron las siguientes situaciones:

- **Actualización de años en el *dataset***: los datos con los que se trabaja en esta entrega son aquellos de 2020 y aquellos disponibles de 2021 por parte de la comisión de limosinas y taxis de Nueva York para trabajar con la información más actualizada en pro del realismo de aplicación. Se identifica que la cantidad de datos disponibles disminuyeron en un 70% a comparación con el 2019, prueba particular del efecto

pandemia. Los datos de días festivos y eventos de tormenta también se actualizan a lo más reciente disponible.

- **Limpieza de *outliers* fuera del percentil 99**: a fin de mantener la mayor cantidad de datos que puedan entregar los mejores resultados, se identificaron aquellos datos que presentaban valores en la variable objetivo por fuera del percentil 99, que presentarían ruido en el modelo. Todos estos son eliminados antes del proceso de enriquecimiento.
- **Creación de nuevas variables objetivo derivadas**: con tal de comprobar si una regresión lineal como en la entrega anterior funciona para el nuevo objetivo de esta entrega, se crean cinco nuevas variables objetivo a partir de la interacción entre la variable objetivo original y los campos predictores esperados. Estas son "amount_per_min", "amount_per_trip", "amount_per_mile", "tip_score" y "demand_score."

4 Selección del modelo

Como se menciona en el enunciado, el conjunto de datos se divide en conjuntos de prueba y entrenamiento por medio de la operación "train_test_split" que se ha visto a lo largo del semestre, para crear nuestras variables X – Y de entrenamiento y de prueba. Se optó que el balance entre ambas fuera de 70%: 30% respectivamente, con una semilla de 101 para recrear a los modelos. Sin embargo, como puede verse a continuación, los resultados por medio de este cálculo son particularmente indeseables, para cada una de las variables generadas.

```
#REGRESIÓN LINEAL
#Modelo para amount_per_min:
X_train, X_test, y_train, y_test = train_test_split(data_dummy, amount_per_min, test_size=0.3, random_state=101)
model_reg1 = LinearRegression()
model_reg1.fit(X_train,y_train)
evaluar_modelo(model_reg1,X_test)
```

```
RMSE: 3216457401.16
MAE: 6255864.64
R²: -23993585786313879552.00
```

```
#Modelo para amount_per_trip:
X_train, X_test, y_train, y_test = train_test_split(data_dummy, amount_per_trip, test_size=0.3, random_state=101)
model_reg2 = LinearRegression()
model_reg2.fit(X_train,y_train)
evaluar_modelo(model_reg2,X_test)
```

```
RMSE: 47616780672.85
MAE: 92612493.46
R²: -18410741121780326400.00
```

```
#Modelo para amount_per_mile:
X_train, X_test, y_train, y_test = train_test_split(data_dummy, amount_per_mile, test_size=0.3, random_state=101)
model_reg3 = LinearRegression()
model_reg3.fit(X_train,y_train)
evaluar_modelo(model_reg3,X_test)
```

```
RMSE: 9182233084.66
MAE: 17859029.85
R²: -1738528994158127308 8.00
```

```
#Modelo para tip_score:
X_train, X_test, y_train, y_test = train_test_split(data_dummy, tip_score, test_size=0.3, random_state=101)
model_reg4 = LinearRegression()
model_reg4.fit(X_train,y_train)
evaluar_modelo(model_reg4,X_test)
```

```
RMSE: 2598351094.45
MAE: 5053676.31
R²: -1860055086134907392.00
```

```
#Modelo para demand_score:
X_train, X_test, y_train, y_test = train_test_split(data_dummy, demand_score, test_size=0.3, random_state=101)
model_reg5 = LinearRegression()
model_reg5.fit(X_train,y_train)
evaluar_modelo(model_reg5,X_test)
```

```
RMSE: 1602673607.87
MAE: 3117128.51
R²: -8056925620430083 84.00
```

Figura 2. Resultados de los modelos de regresión lineal

Se identifica que la forma en que se plantea el *dataset* lo dota de una particular codependencia en las columnas, debido a la manera en que se obtienen estas columnas para empezar.

5 Evaluación y despliegue del modelo

Con base en lo anterior, se realiza la selección de modelos de Ridge y Lasso por sus características de lidiar con codependencia de columnas, a la vez que se utiliza *Grid Search*

para identificar los mejores parámetros para ello. Los resultados, vistos a continuación, fueron iguales para ambos. De entre las suposiciones que se hacen en la entrega, está cómo se espera que el hecho de un día feriado o un día particularmente tormentoso aumente las posibles ganancias del taxista. La mejoría es evidente, aunque podría tener amplia mejoría.

```
RMSE: 0.60
MAE: 0.26
R²: 0.14
```

Figura 3. Resultados de los modelos de Ridge y Lasso

Al pensar en posibilidades de mejora, se destaca la búsqueda de mejores parámetros, el correcto balanceo de los datos en preprocesamiento, y la búsqueda de otros modelos que puedan ajustarse mejor a los datos.

6 Cierre de *sprint*

Después de lo que se consigue con este *sprint*, esta sección representa las consideraciones de cierre de esta entrega.

6.1 Terceras conclusiones

La limpieza de los datos es un proceso que requiere de muchas más iteraciones de las que se pensaba al inicio de semestre. Todo el valor que puede derivarse de este quehacer proviene de los datos; si éstos no son de calidad, no puede producirse algo de calidad. Puede que sea un proceso crítico, pero lo más destacable es la realidad de cómo los datos pueden estar en condiciones particularmente paralizantes a la hora de continuar con algún desarrollo.

Los tiempos de ejecución no pueden subestimarse; al trabajar con volúmenes que alcanzan a calificar como Big Data pueden ser prohibitivos en término de qué algoritmos y modelos pueden aplicarse fácilmente debido a los tiempos de ejecución y su propia complejidad.

6.2 Acciones sugeridas

Priorizar la búsqueda de modelos y algoritmos de baja complejidad algorítmica, a fin de maximizar las posibles soluciones que pueden plantearse para situaciones como esta, donde se requiere de un enorme volumen de datos.

Considerar la posibilidad de continuar con este proyecto, a fin de planear una herramienta que verdaderamente sea útil en su totalidad para los conductores de taxi de Nueva York.

Dividir mejor las tareas, de forma que se cumpla con más del proyecto, como el desarrollo REST, en vez de intentar conseguir un resultado de distintos modelos que no dan en el tiempo necesario.

6.3 Contribución del equipo

El desarrollo del taller se realizó en trabajo conjunto entre los miembros del equipo, con ciertas personas liderando alguna de las fases en exclusivo. Jose Fernando Lideró la fase de limpieza y análisis de datos, Felipe Gutiérrez lideró la fase de desarrollo de los modelos de regresión lineal, y Alejandro Mantilla lideró la fase de desarrollo de los modelos Ridge y Lasso.

7 Bibliografía

1. **NYC Taxi & Limousine Commission.** ***Taxi Zona Maps and Lookup Tables.*** **[En línea] [Citado el: 13 de noviembre de 2021.] https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page.**

2. NYC OpenData. ***Bus Breakdowns and Delays.*** **[En línea] [Citado el: 13 de noviembre de 2021.] https://data.cityofnewyork.us/Transportation/Bus-Breakdown-and-Delays/ez4e-fazm.**

3. GeeksforGeeks. ***Decision Trees.*** **[En línea] [Citado el: 15 de noviembre de 2021.] https://www.geeksforgeeks.org/decision-tree/.**

4. Public Holidays Global. ***Start Planning Nueva York.*** **[En línea] [Citado el: 15 de noviembre de 2021.] https://publicholidays.com/us/es/new-york/2019-dates/.**

5. National Center For Environmental Information. ***Storm Event Database.*** **[En línea] [Citado el: 15 de noviembre de 2021.] https://www.ncdc.noaa.gov/stormevents/details.jsp.**