Exploratory Data Analysis of New York City TLC Data

Executive Summary

Prepared by Automatidata

The NYC Taxi & Limousine Commission has contracted with Automatidata to build a regression model that predicts taxi cab ride fares. In this part of the project, the data needs to be processed in the next steps prior to any modeling: analyze, explore, clean and structure.

Key Insights

Problems:

After an initial exploration of the data, we found inconsistencies that could be an obstacle in performing the model. First, we explored the relationship between distance and the established fee and found costs in trips with 0 distance. On the other hand variables such as in number of passengers (also with 0 values), fees (negative fees exist) and tolls (negative tolls exist) could be a problem.

Possible solutions:

We propose to eliminate in a first instance the trips with 0 distance and those with negative costs. To avoid problems in the prediction of quotas in the model

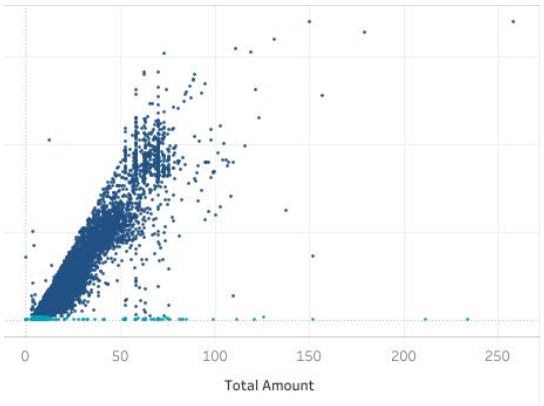
Set negative tolls as 0 and adjust costs with the changes.

Key Points:

Confirm data provided by NYC TLC and obtain available metadata.

Establish the most significant variables affecting travel costs.

Details



In the results obtained in the exploration of the data. The Automatidata team contrasted the trip distance variables and the final trip costs. As can be seen in the light blue dots. There are a large number of trips with 0 distance and high costs.

Next Steps

- Identify and filter data with discrepancies, such as those shown in the graph, for modeling.
- Determine the most relevant variables for predicting travel costs.
- Perform statistical tests and hypothesis development with the variables chosen for model development.

Tik Tok Claim Classification Project
Exploratory Data Analysis (EDA)

Executive Summary

Prepared by TIk Tok Data Team

The TikTok data team seeks to develop a machine learning model to assist in the classification of claims for user submissions. In this part of the project, the data needs to be analyzed, explored, cleaned, and structured prior to any model building.

Key Insights

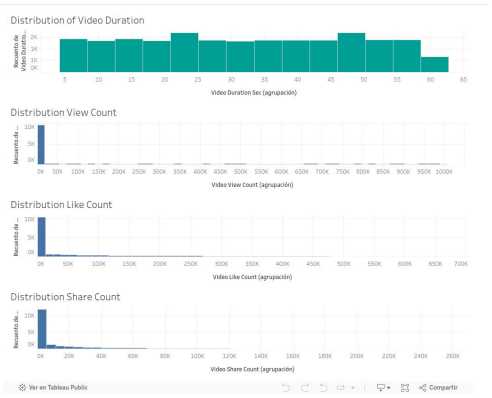
Problem: The TikTok data team conducted exploratory data analysis at this stage. The purpose of the exploratory data analysis was to understand the impact that videos have on TikTok users. To do so, the TikTok data team analyzed variables that would showcase user engagement: view, like, and comment count.

From the findings found in the analysis, we found relevant differences between a claim and an opinion on the variables of viewing and likes on a video.

Key insights

- Engagement seems to be a useful variable for the prediction between claim and opinion.
- Claim and opinion videos that have a higher engagement than the rest tend to come from banned or future banned authors and could be a good predictor to determine this status in new video authors.

Details



The key component of this project's exploratory data analysis involves visualizing the data. As illustrated in the following histograms, it is clear that the vast majority of videos are grouped at the bottom of the range of values for three variables that showcase TikTok users (video viewers') engagement with the videos included in this dataset.

Next Steps

The analysis aims to take into account two main problems with the data when building the model.

Null values
The existence of null values in the data must be taken into account in the construction of the model. For this reason it is necessary to understand the null values in the data.

Skewed data distribution
Video view and like counts are all concentrated on low end of 1,000 for opinions. Therefore, the data distribution is right-skewed, which will inform the models and model types that will be built.

Waze Churn Prediction Project
Exploratory Data Analysis (EDA)

Executive Summary

Prepared by Waze Leadership Team

The Waze data team is currently developing a data analytics project aimed at increasing overall growth by preventing monthly user churn on the Waze app. Thorough exploratory data analysis (EDA) enables Waze to make better decisions about how to proactively target users likely to churn, thereby improving retention and overall customer satisfaction.

Key Insights

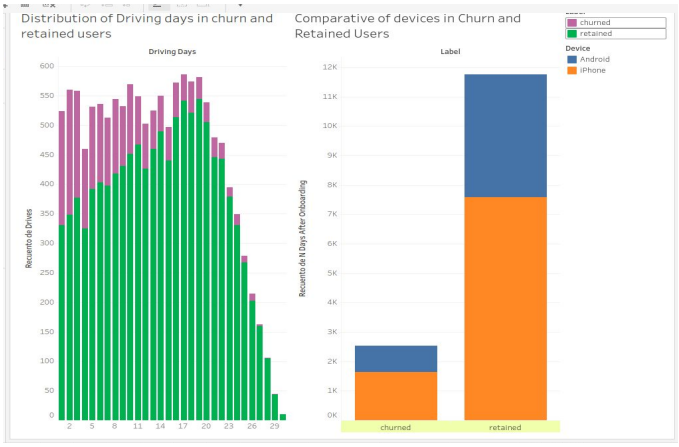
Consistent use of the app decreases the probability of user churn. It was found that 40% of users who churned did not use the app during the month. Meanwhile, none of the users who used the app for the entire month did so.

The distributions that are present in the variables are two. Variables with a uniform distribution, and variables with a skewed distribution to the right where their highest frequency is at the lowest values.

There are problems with the data reflected in the variables. This is because the data reflected in the variables “kilometers driven”, “Time driven” are highly skewed. The data reflected in the variables “kilometers driven”, “time driven” are highly improbable and sometimes inconsistent with other data, such as “day of activity” or “number of times driven”.

Finally, during the exploration conducted, a tendency was found in churn users to perform multiple sessions in a day driving a lot of time in a few days. Research and data collection is needed to understand this finding.

Details



The churn rate is highest for people who didn't use Waze much during the last month.

The proportion of churned users to retained users is consistent between device types.

Next Steps

- ➔ Investigate the erroneous or problematic discrepancies between number of sessions, driving_days, and activity_days.
- ➔ Continue to explore user profiles with the greater Waze team; this may glean insights on the reason for the long distance drivers’ churn rate.
- ➔ Plan to run deeper statistical analyses on the variables in the data to determine their impact on user churn.