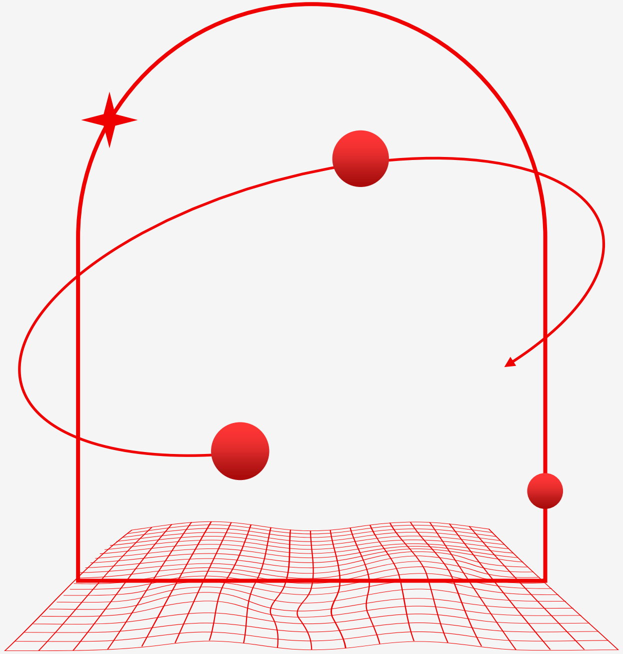


Rappi Experimentation & Analitics Senior Case

Whoa!

Juan Carranza
jdcarranzas@outlook.com

Table of contents

01 **Bussiness Case**
What are we looking for?

02 **EDA**
What we learned from the data

03 **Prior Analysis**
Possible patterns in the data

04 **Experiment**
How we set the environment

05 **Modelling**
Techniques and results

06 **Recomendations**
Final conclusions and remarks

01

Bussiness Case

What are we looking for?

Rappi

Introduction

Rappi’s Operations team is interested in decreasing the number of orders that are not taken by any courier, due to the fact that they are not attractive enough for couriers.

Scope

Medition

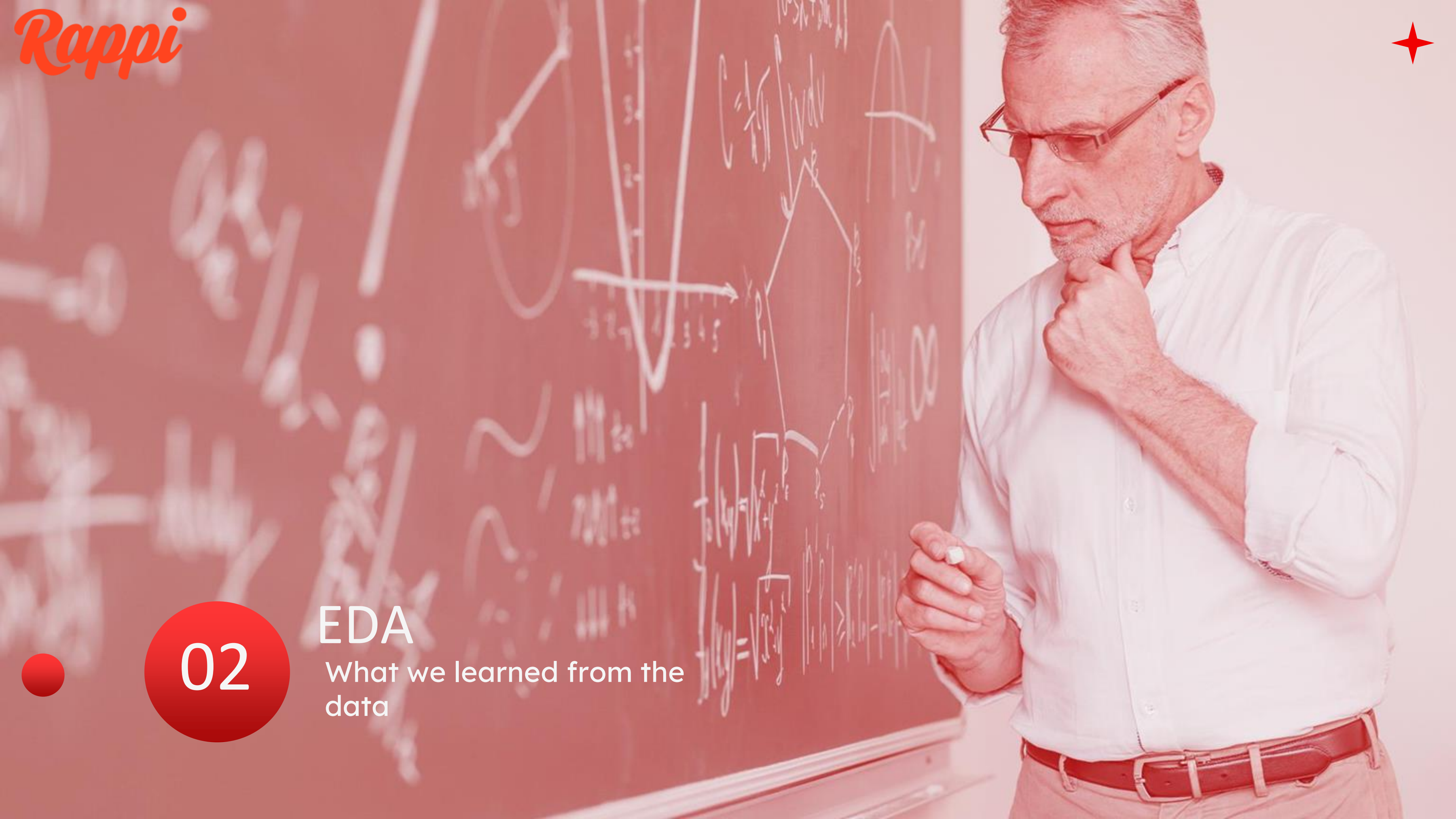
Date time data, measured in one month. September 2017.

Variables

Distance from user to store (km), difference in meters between the store and user altitude, total earning, taken as a binary variable: 1 if taken, 0 otherwise.

Objective

Identify key drivers that might predict if a given order is taken or not.



Rappi

02

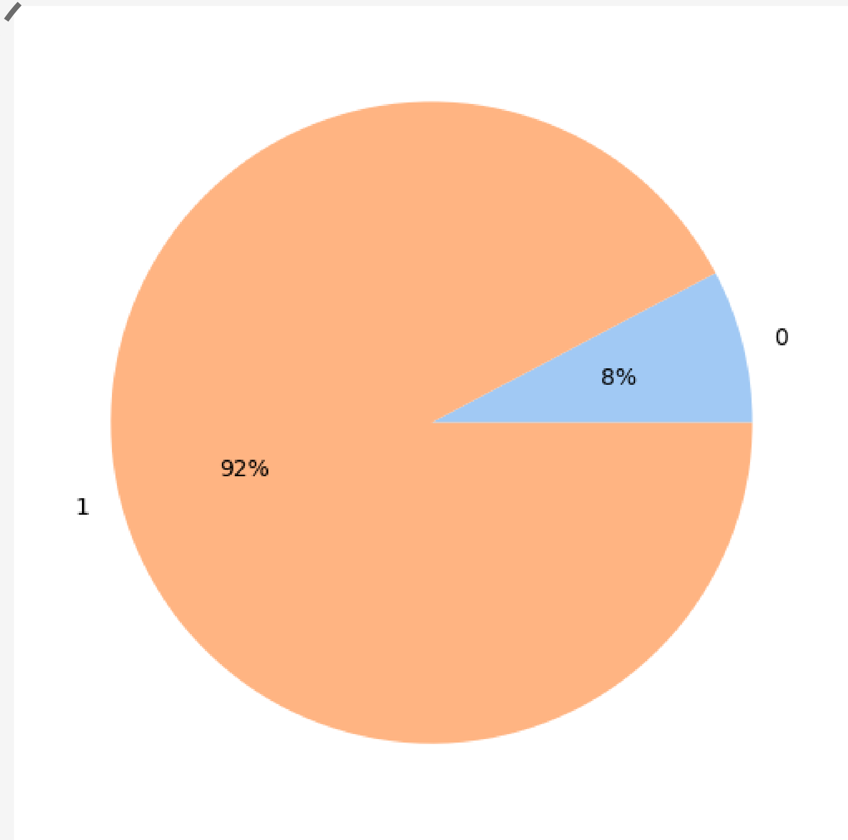
EDA

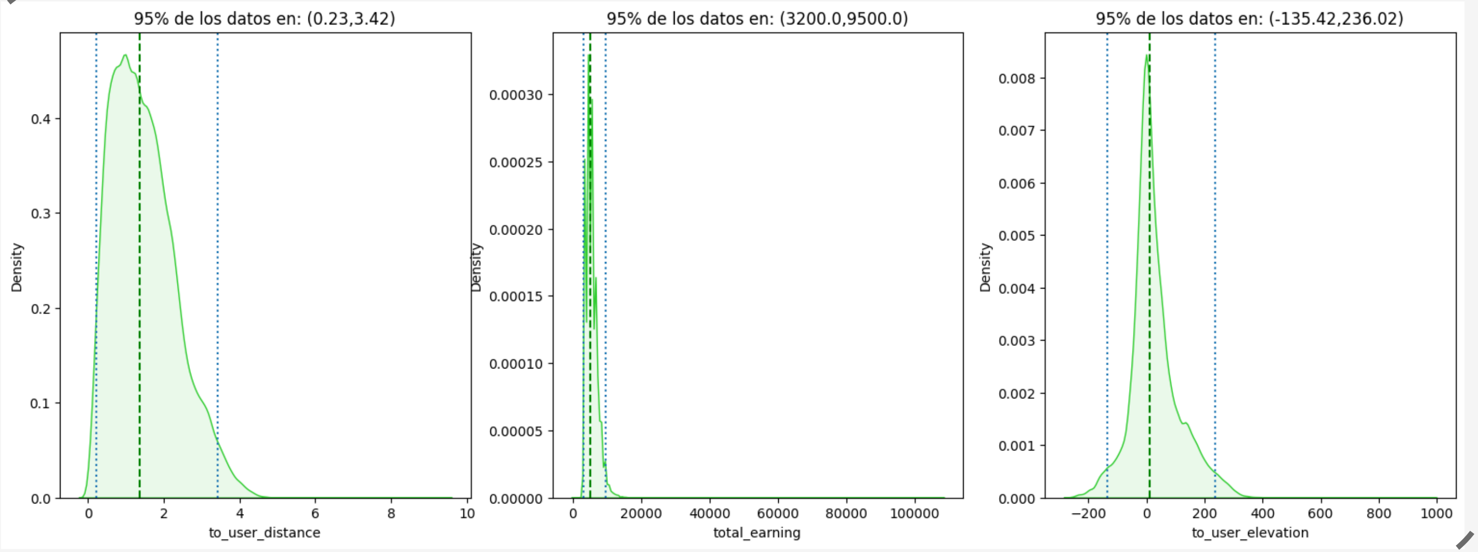
What we learned from the data

Percentage of non taken orders

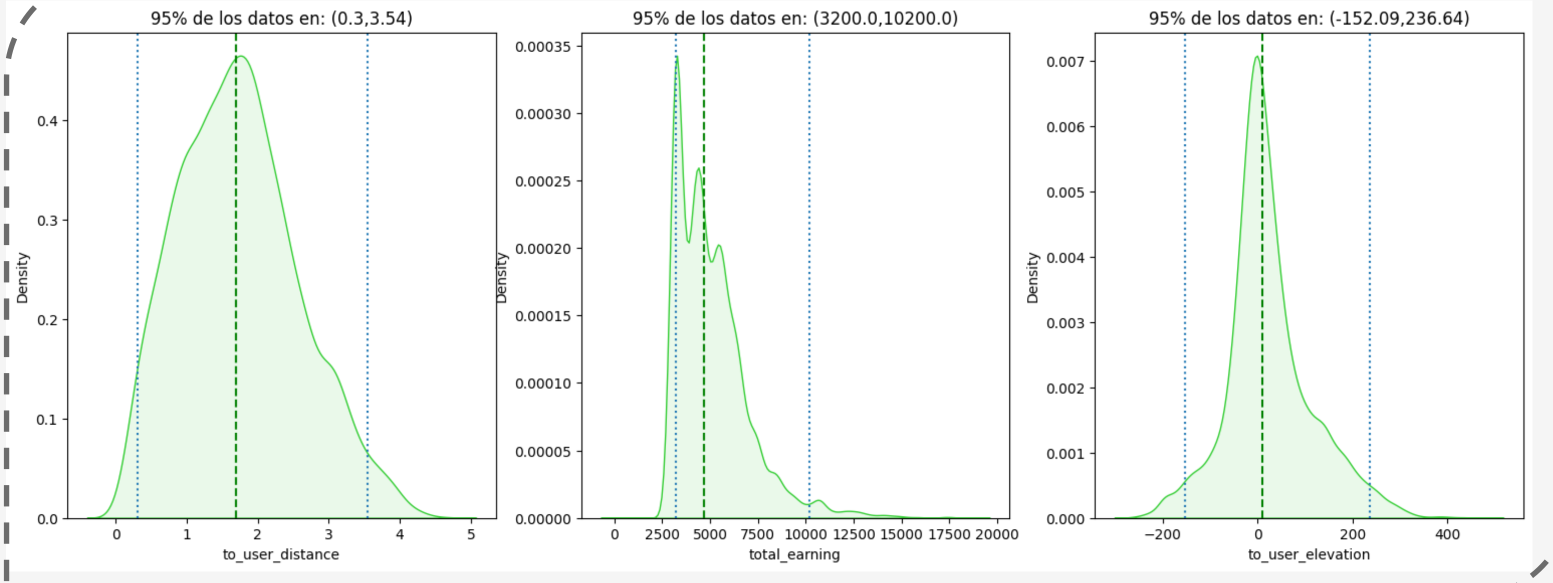
92% of the orders were taken by any Courier, 8% were not taken by any Courier.

Taken (1)	Not taken (0)
115860	9689

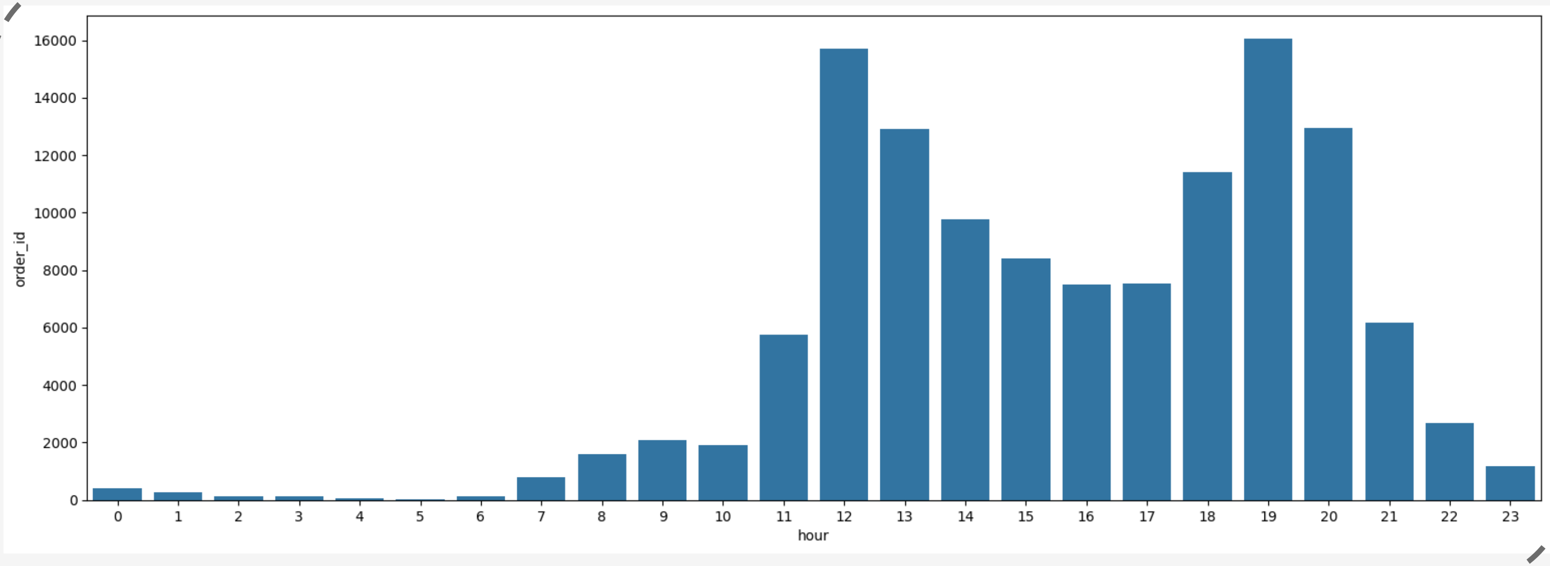




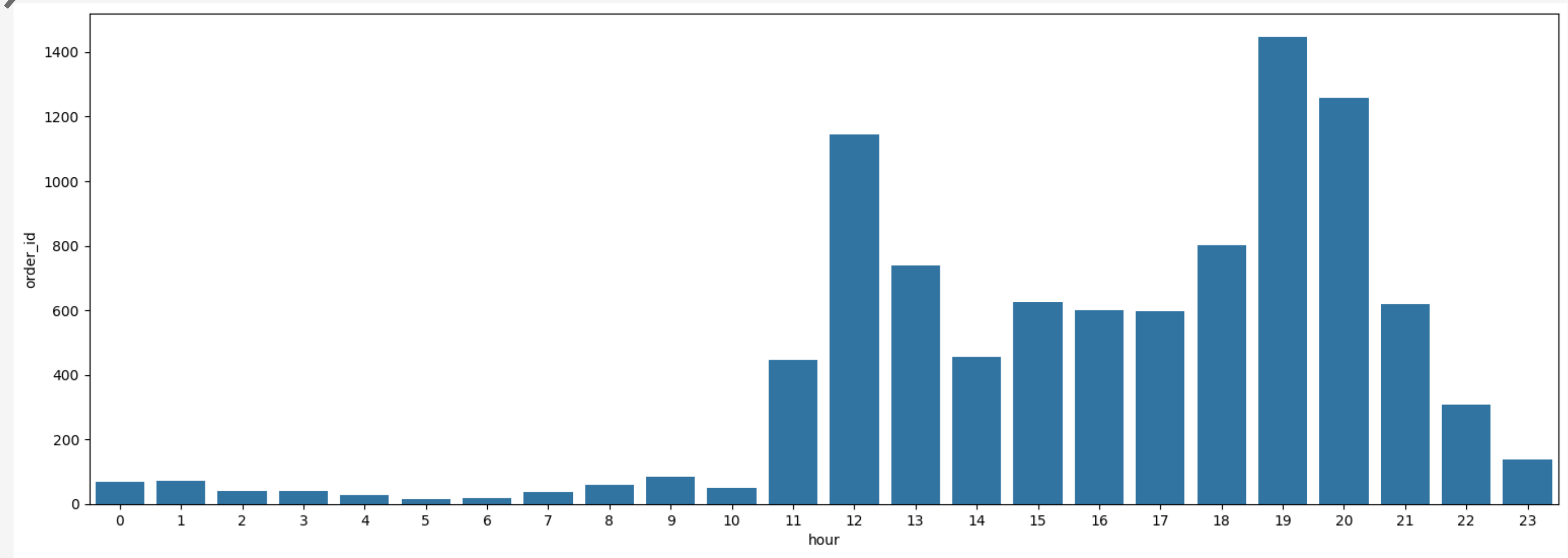
95% of the clients are between 0.23km and 3.42km away from the store.
95% of total earnings of a Courier is between $3200 and $9500
95% of user elevation to the store is between -135.42m and 236.02m



Checking only non taken orders:
95% of the clients are between 0.3km and 3.54km away from the store.
95% of total earnings of a Courier is between $3200 and $10500
95% of user elevation to the store is between -152.09m and 236.64m



The peak hours in terms of orders created usually matches the lunch time and the dinner time, with steady levels in the afternoon and the early night (around 20 to 21 hours).



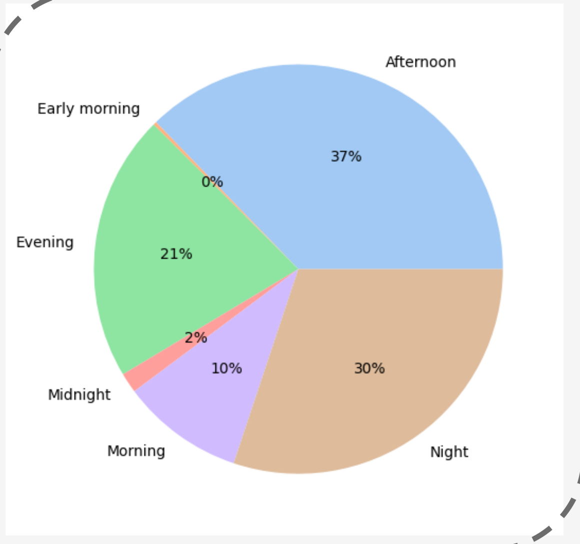
Checking only non taken orders, we have critical peaks at the lunchtime and most of the non taken order are concentrated in the late-afternoon – night hours.

Categorizing the data by hour

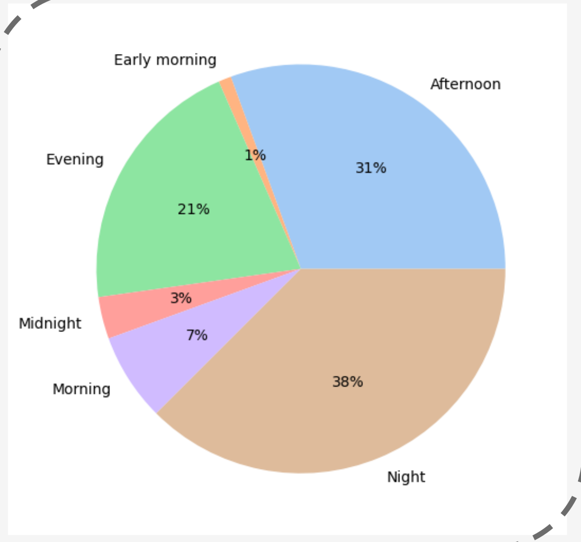
If we categorize the hour of the day when the order was originally created, we can gain some general insights among taken and non taken orders.

Category	Hours
Early Morning	3 – 6
Morning	7 – 11
Afternoon	12 – 15
Evening	16 – 18
Night	19 – 22
Midnight	23 - 2

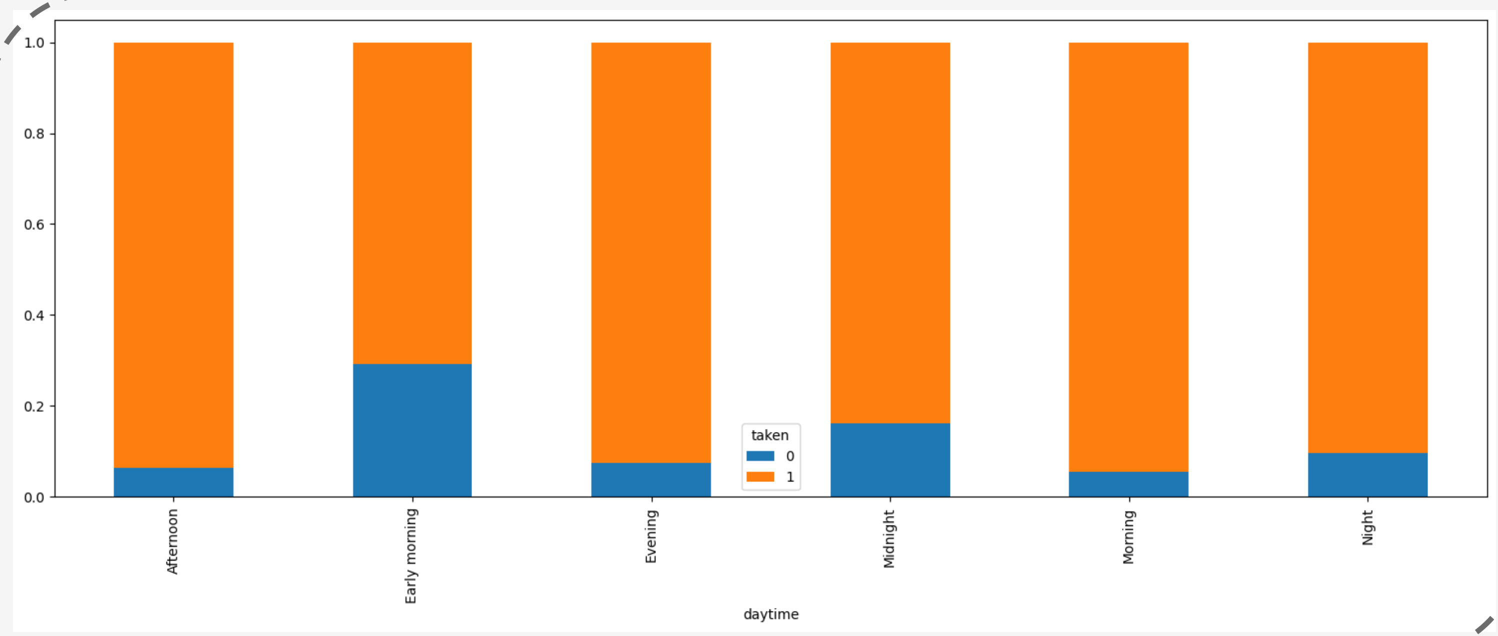
Total orders



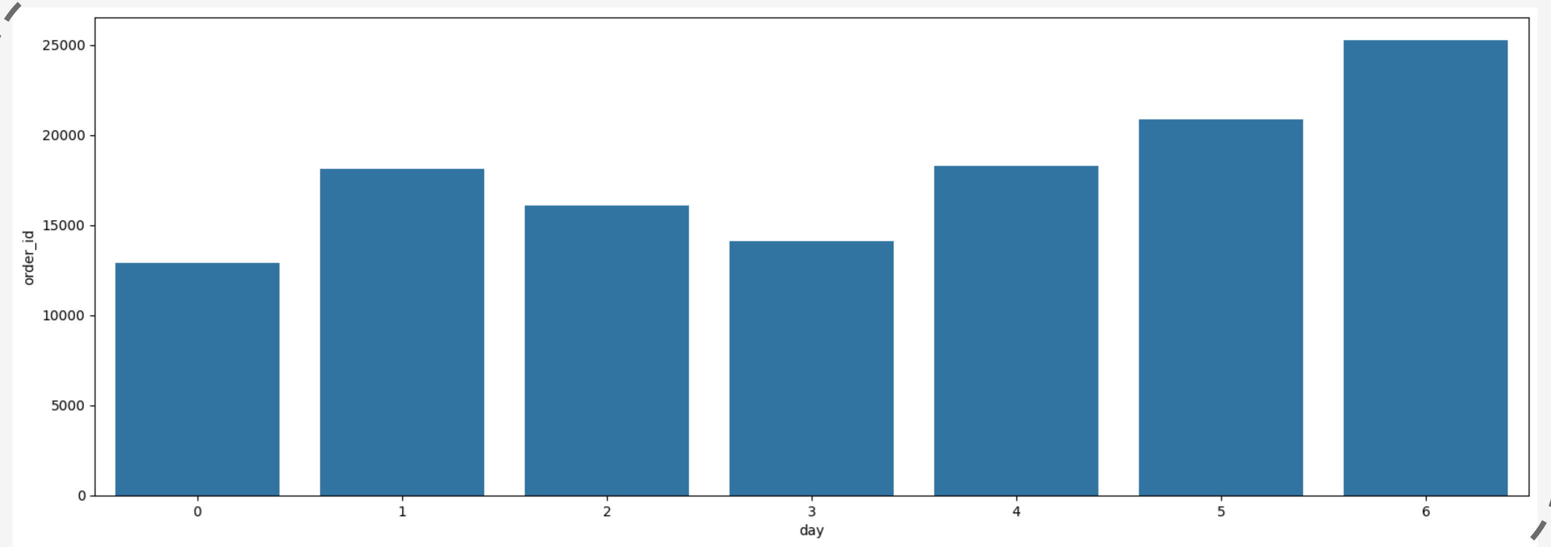
Non taken orders



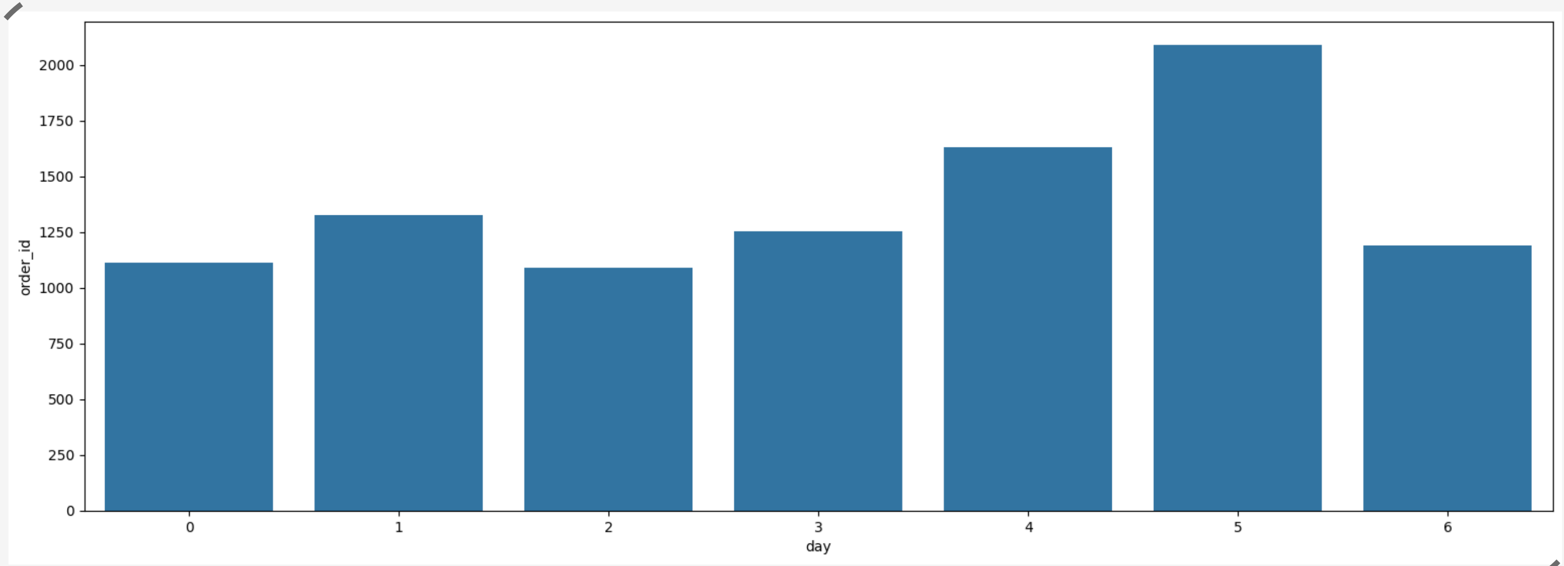
The highest percentage of non taken orders are in the night



But relatively, in the early morning and in the midnight is more probable that an order may not be accepted by any Courier.



The days that have most orders are on the weekends (Friday, Saturday and Sunday) and the Tuesday.



But the days with more non taken orders are the Fridays and Saturdays. The day with less non taken orders is Sunday.

Prior Analysis

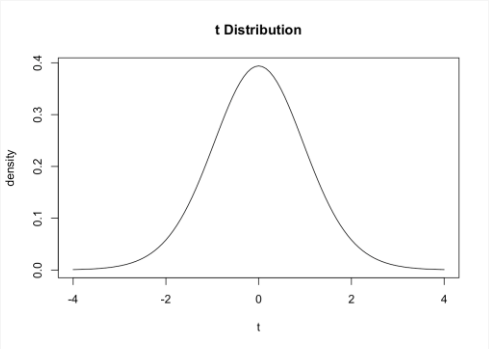
Possible patterns in the data

03

T-test

We test for mean differences with the null hypothesis that the taken orders have lesser values that the non taken orders.

We assume different variances for each group for robustness.



T-test for difference of means

The total earnings is higher in the ‘taken’ group.
P-value = 2.3e-55

The total distance is lower in the ‘taken’ group.
P-value = 1

The difference in altitude is positive and higher in the ‘taken’ group
P-value = 1.1e-5

04

Experiment

How we set the environment

Solving problems

Unbalanced Dataset

Undersampling for majority class (taken orders)

Normalizing

For continuous variables, in order to solve high variance

One hot encoding

For considering discrete variables (day, hour)

Cross-Validation

In order to assure the quality of the experiment

05

Modelling

Techniques and results

Rappi

Modelling Results

6 Types of classification models tested: 3 of them ensembles, logistic regression, Naïve Bayes, SVC.

33 Different configurations of parameters, with 3 validation folds.

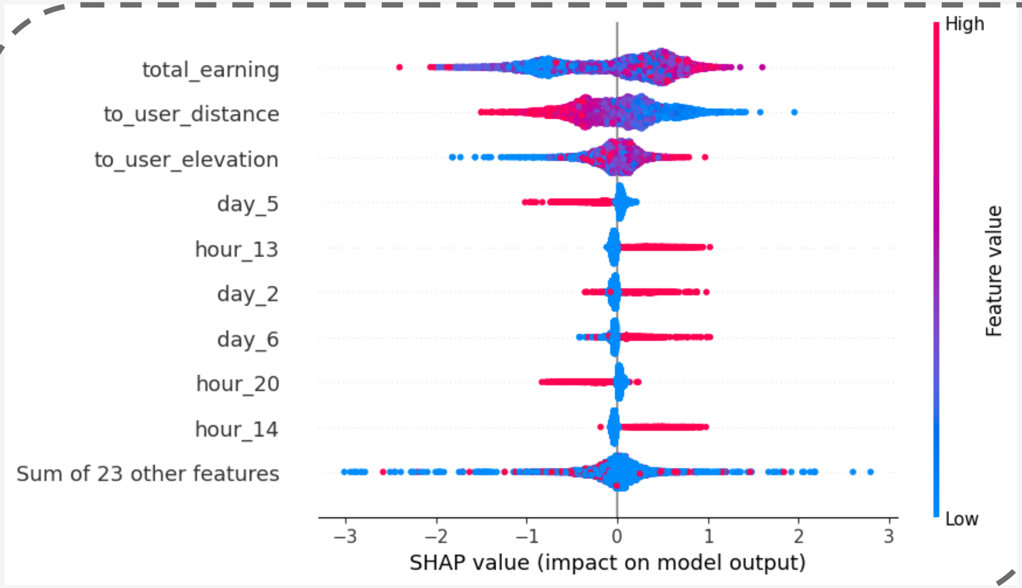
5 Different metrics of classification: accuracy, f1, precision, recall and ROC - AUC

8 Highly relevant variables capable of giving valuable insights

XGBoost

Got the highest score! 0.66 average in the train and test split across all metrics!

Variable Interpretation



- SHAP values strongly suggest that the 20 hour and the Saturday are strong predictors whenever an order is not accepted. Higher total earning and lower user distance are strong predictors of higher acceptance.

06

Recommendations

Final conclusions and remarks

Conclusions

Key Hours

Offer higher rates at night to the couriers

Rates for distance

Offer higher rates if the distance is higher than 2.8km

Happy Weekdays

Offer incentives to pick more orders on saturdays

Lunch hours

Encourage the user to order before peak hours to avoid saturation in the lunch and dinner hour

Weather checking

For an upcoming challenge, take into account the weather information

Thank you!