CHURN is Coming!

House G58

Overview

Data Cleaning

EDA

Model Selection

Performance

Solutions

Impact / Thought

Clean up

Dates: strings to Datetime type

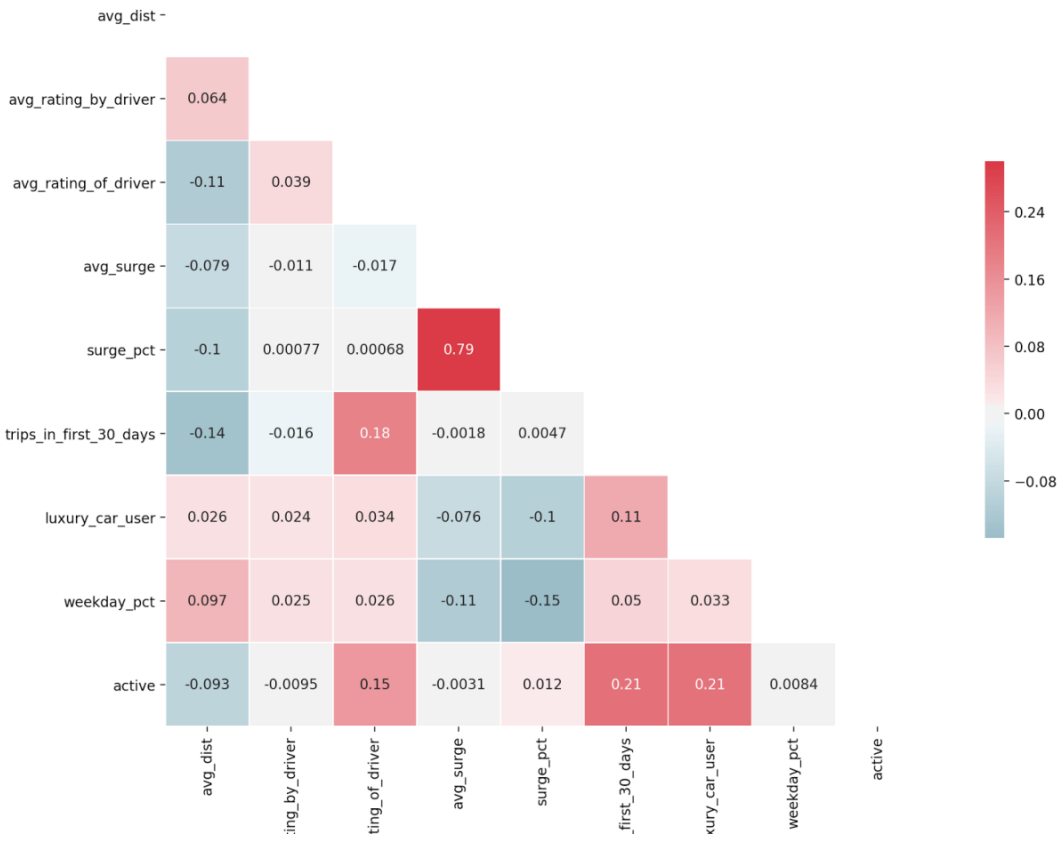
City and Phone to Dummies

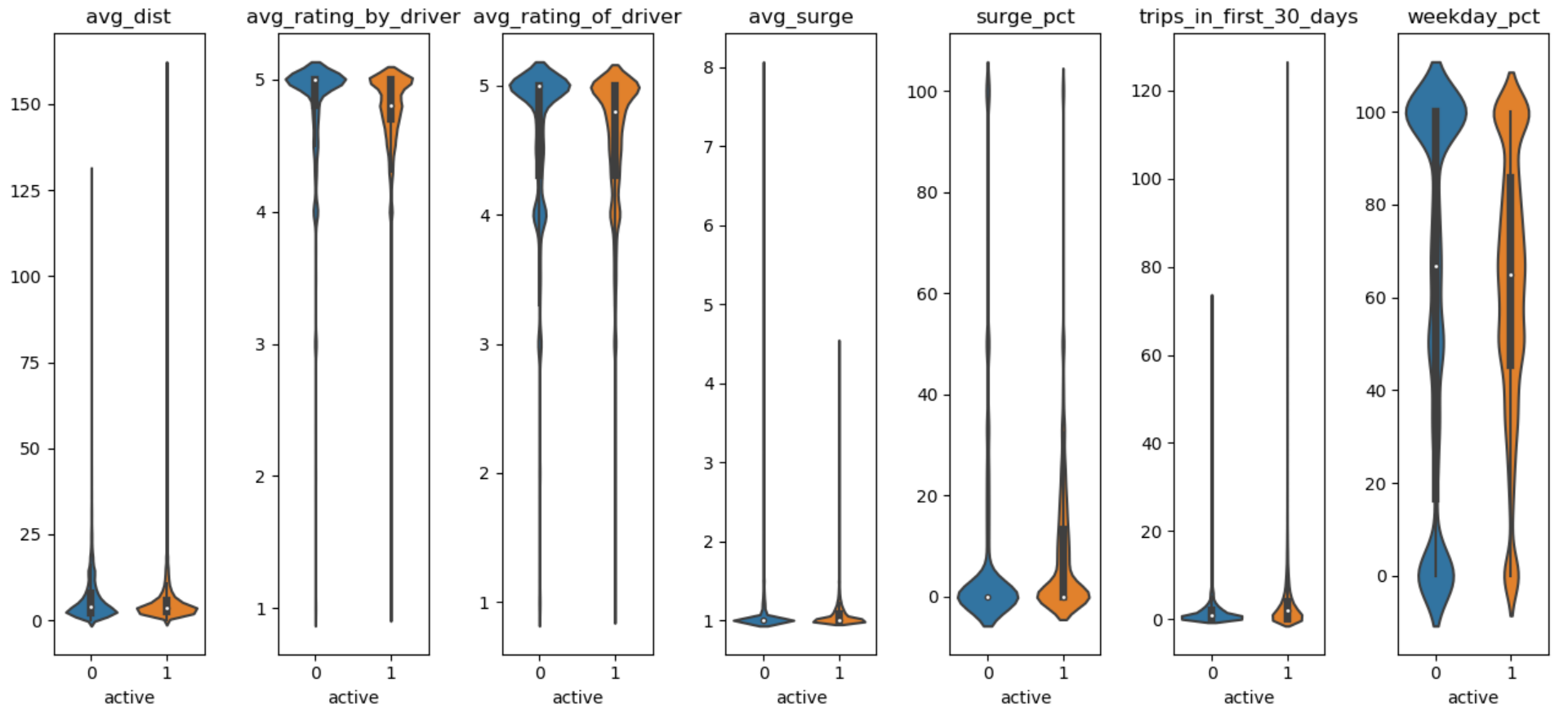
Changed NaNs - MICE imputation

Created an active column for the dependent variable

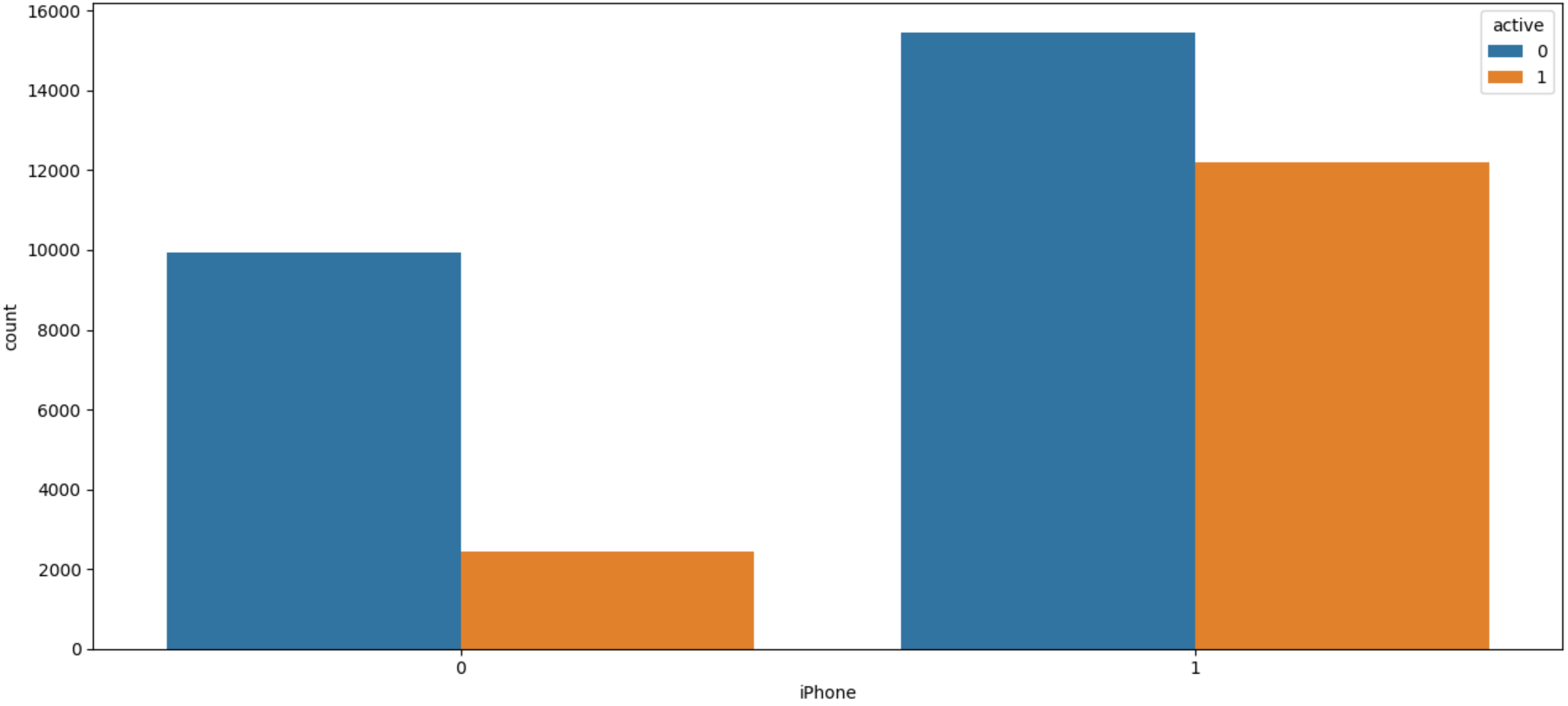
Dropped Dates

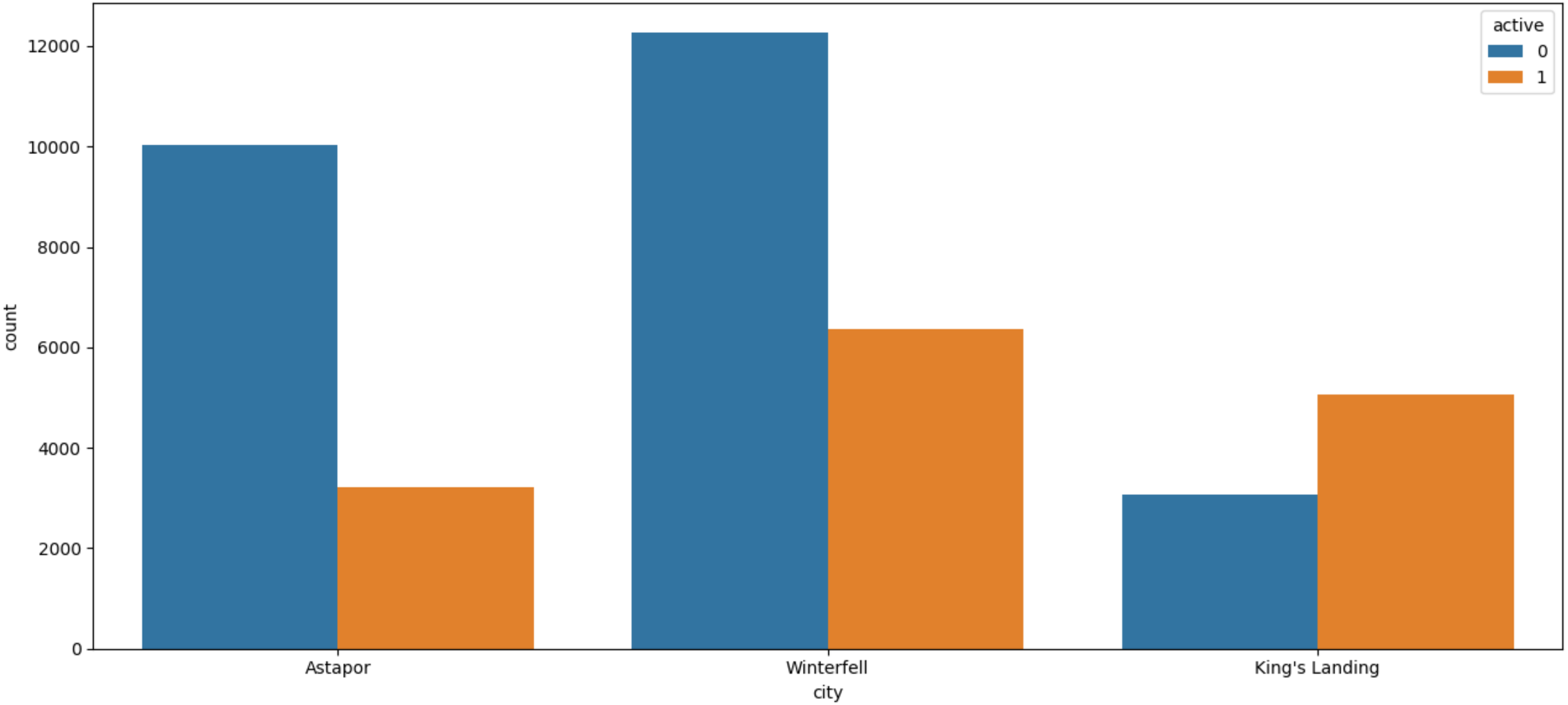
Ice Map

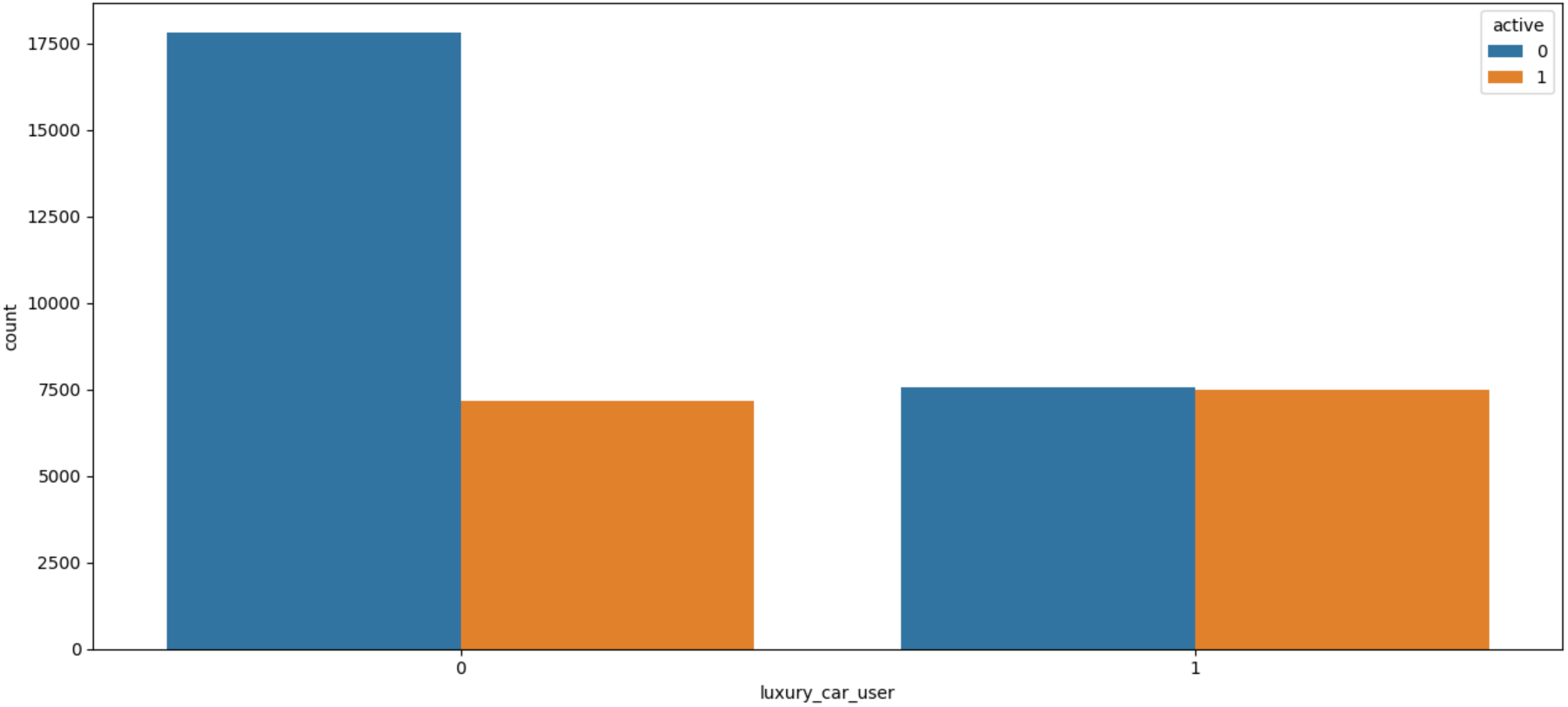

avg_dist
avg_rating_by_driver
avg_rating_of_driver
avg_surge
surge_pct
trips_in_first_30_days
luxury_car_user
weekday_pct
active
0.064
-0.11 0.039
-0.079 -0.011 -0.017
-0.1 0.00077 0.00068 0.79
-0.14 -0.016 0.18 -0.0018 0.0047
0.026 0.024 0.034 -0.076 -0.1 0.11
0.097 0.025 0.026 -0.11 -0.15 0.05 0.033
-0.093 -0.0095 0.15 -0.0031 0.012 0.21 0.21 0.0084
0.24
0.16
0.08
0.00
−0.08

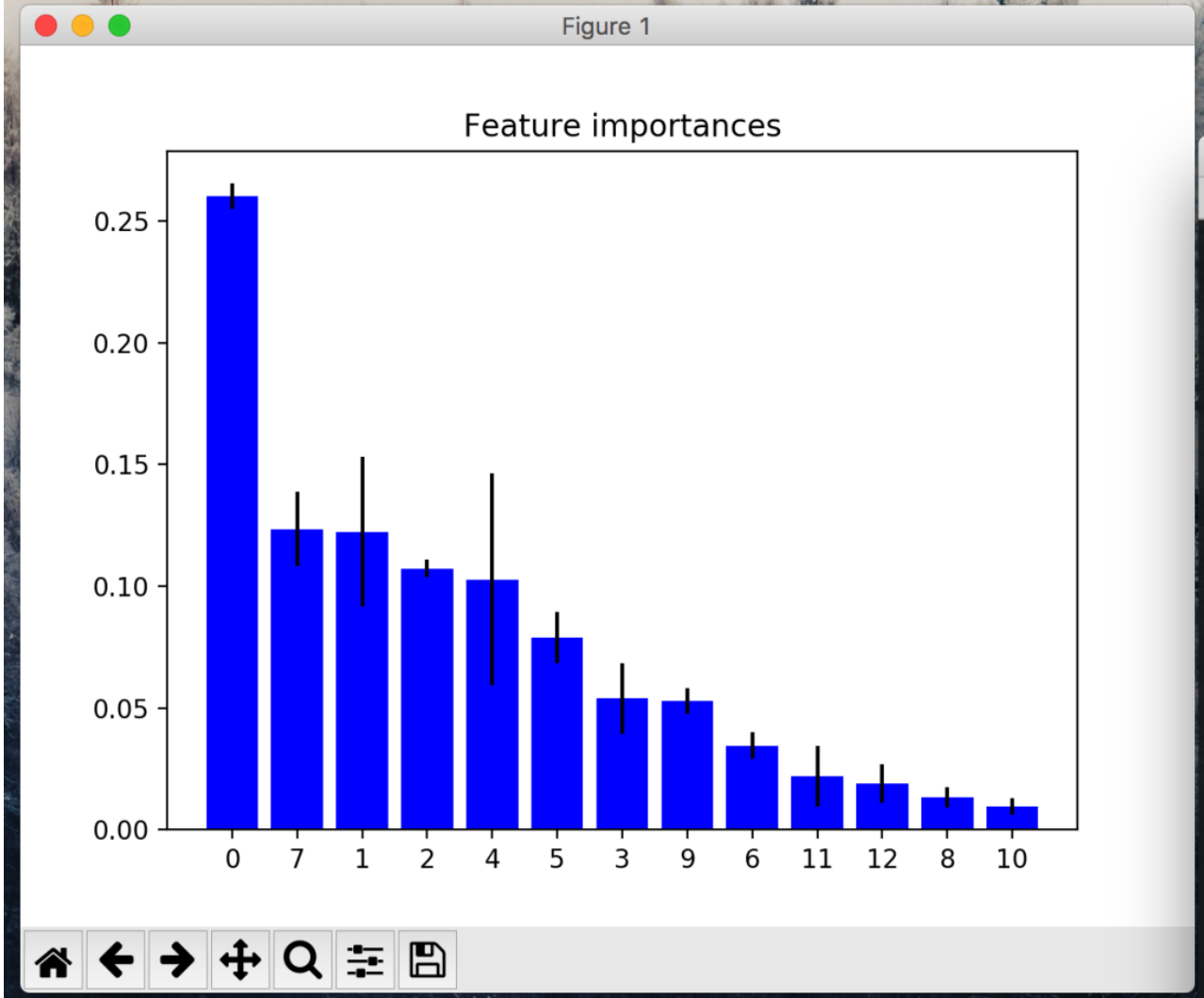


Pre standardization


16000
14000
12000
10000
8000
6000
4000
2000
0
count
0
1
iPhone
active
0
1


active
0
1
12000
10000
8000
6000
4000
2000
0
count
Astapor
Winterfell
King's Landing
city


active
0
1
17500
15000
12500
10000
7500
5000
2500
0
count
0
1
luxury_car_user



IPython: tue/case_st... ⌘1 IPython: galva

```
In [198]: run feature_importances.py
Feature ranking:

1 : avg_dist - 0.260
2 : signup_date - 0.123
3 : avg_rating_by_driver - 0.122
4 : avg_rating_of_driver - 0.107
5 : city - 0.103
6 : last_trip_date - 0.079
7 : avg_surge - 0.054
8 : trips_in_first_30_days - 0.053
9 : phone - 0.034
10 : weekday_pct - 0.022
11 : active - 0.019
12 : surge_pct - 0.013
13 : luxury_car_user - 0.010
```

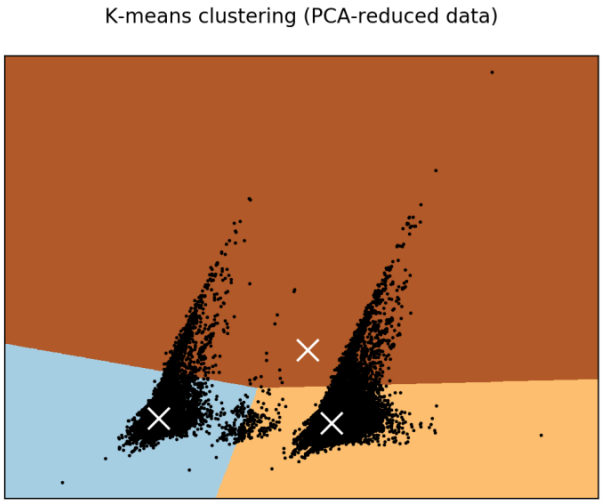
Model Selection
01

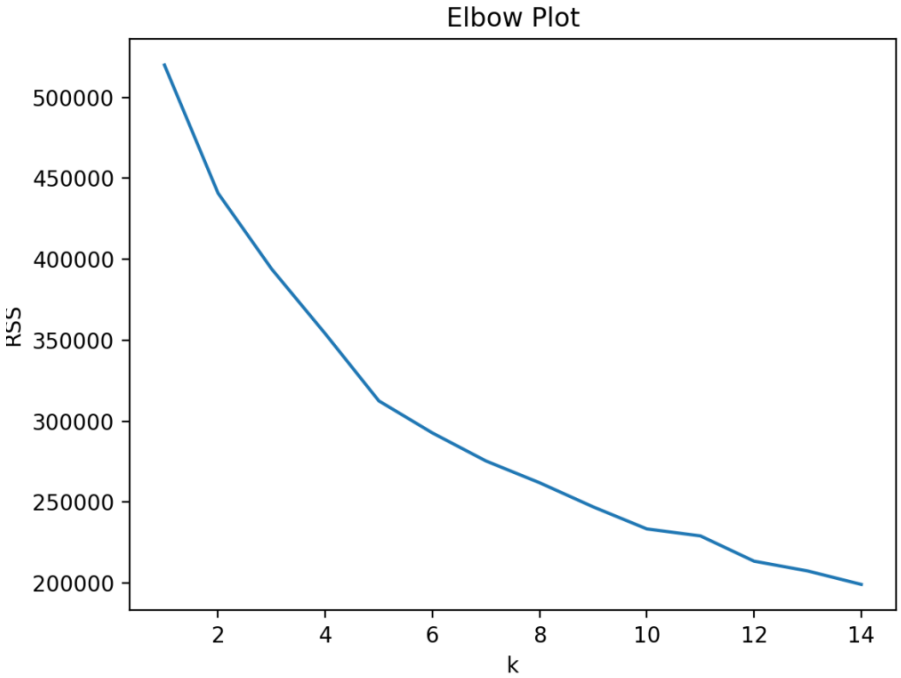
Gradient Boost Classifier
.79
XGBoost Classifier
.79
LGBM Classifier
.78

	CVMean	Std
Logistic Regression	0.720300	0.005745
KNN	0.759850	0.007661
Decision Tree	0.703300	0.002720
Naive Bayes	0.702200	0.007098
Random Forest	0.749675	0.004617

Model Selection

02





Model Selection 3
XGBoost

XGBClassifier(colsample_bylevel=1, colsample_bytree=0.55, gamma=3, learning_rate=0.05, max_delta_step=0, max_depth=3, min_child_weight=1.5, missing=None, n_estimators=2200, nthread =-1,objective='binary:logistic', reg_alpha=0.464, reg_lambda= 0.8571,scale_pos_weight=1, seed=5, silent=1, subsample=0.5213)

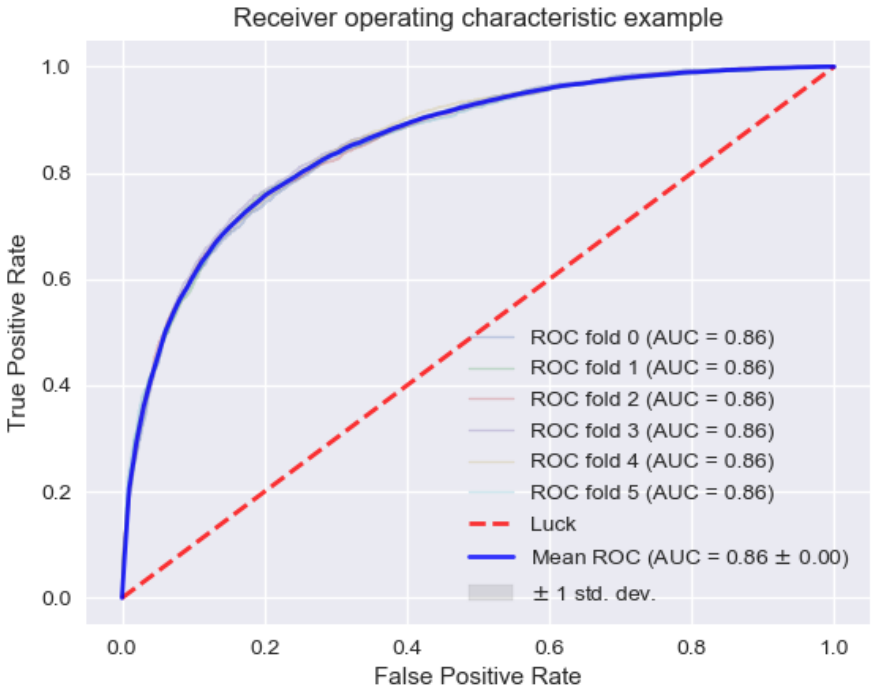
Performance

Accuracy = 7907 / 10000 = 79.5%

```
[[5414  800]
 [1293 2493]]
```

Recall = 2493 / (2493 + 1293) = 66%

Precision = 2493 / (2493 + 800) = 76%



Problems to solve

1. Leakage - after making the active users label, drop ‘last trip’ feature

2. Cost Benefit?

Proposed solution

Market to Androids

Use Incentives

Cost benefit matrix on these solutions

Thoughts

- Drop one of the surge columns

- Print precision and recall while tuning parameters

Thank you.