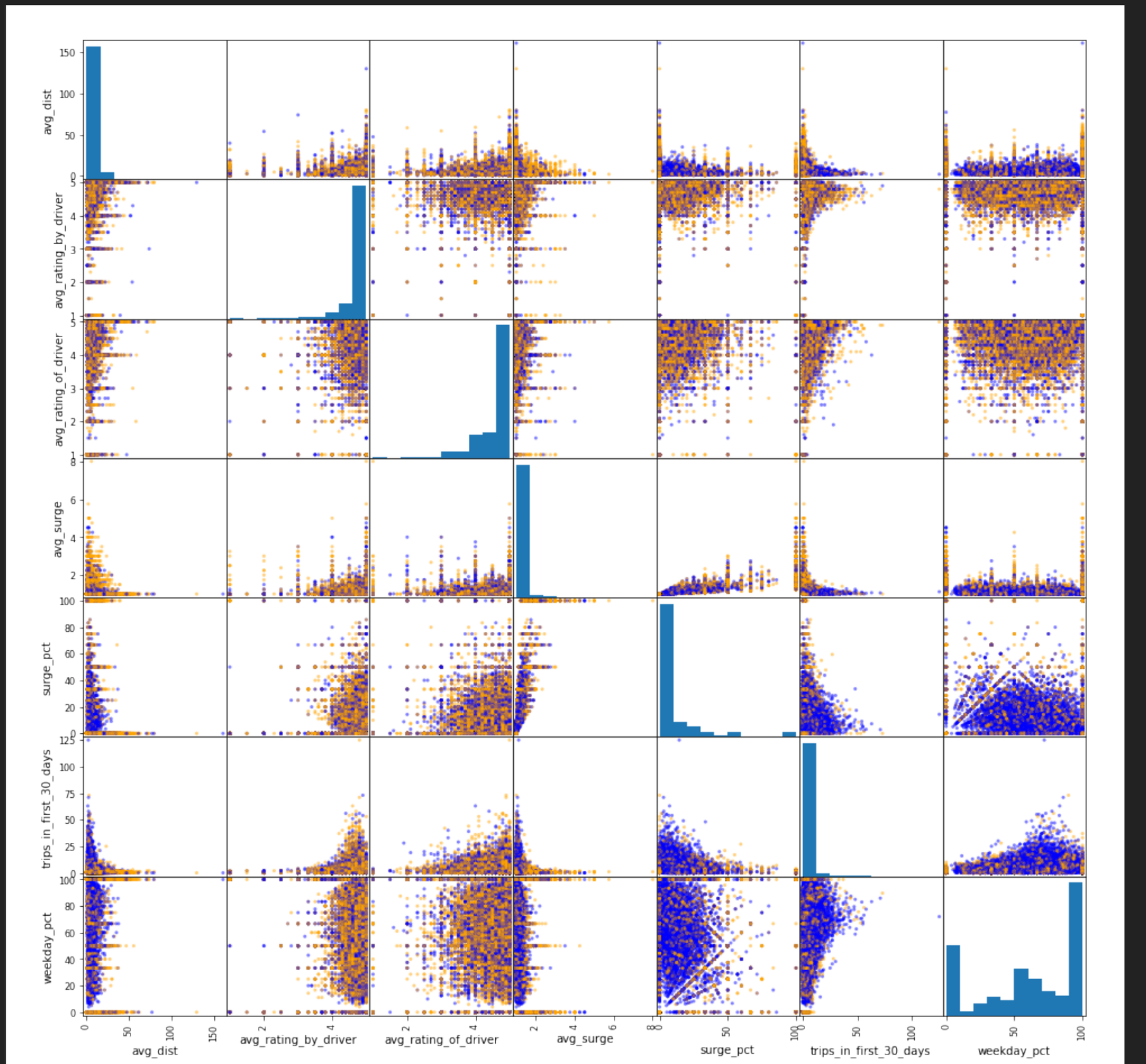
CHURN PREDICTION CASE STUDY

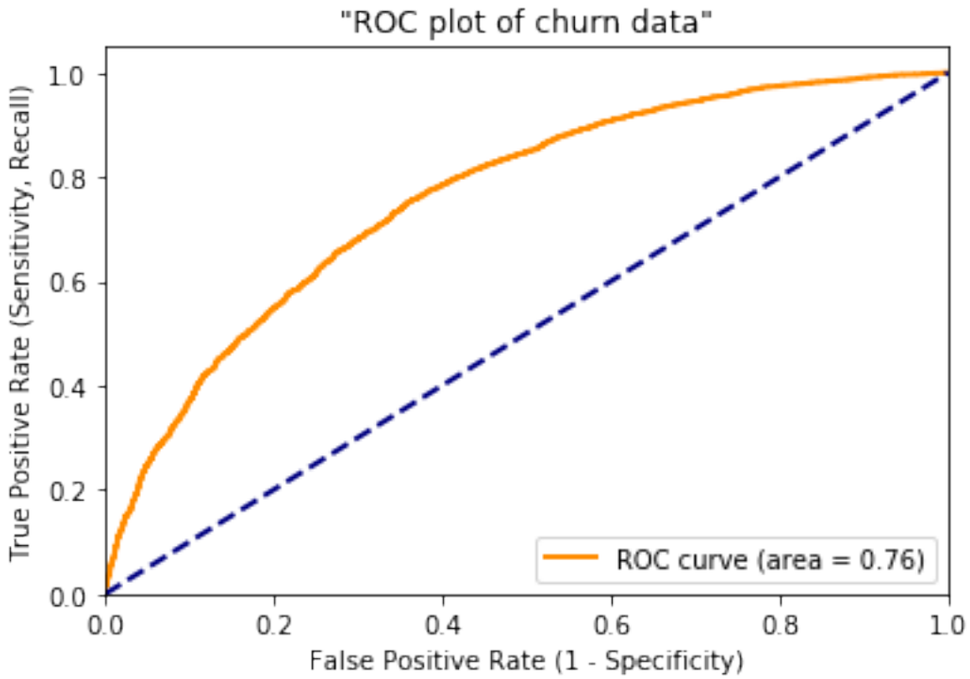
EDA OBSERVATIONS

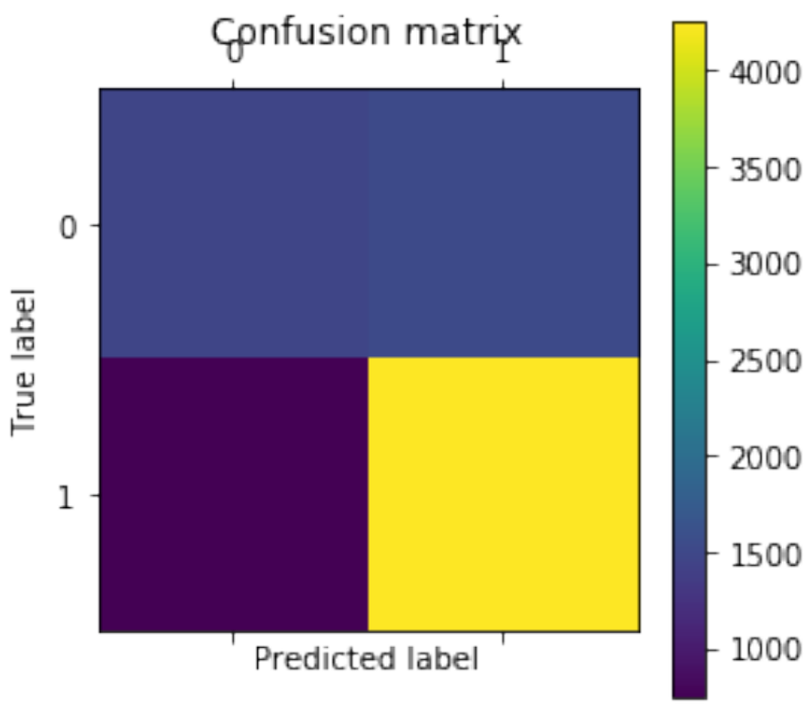


SCATTER MATRIX

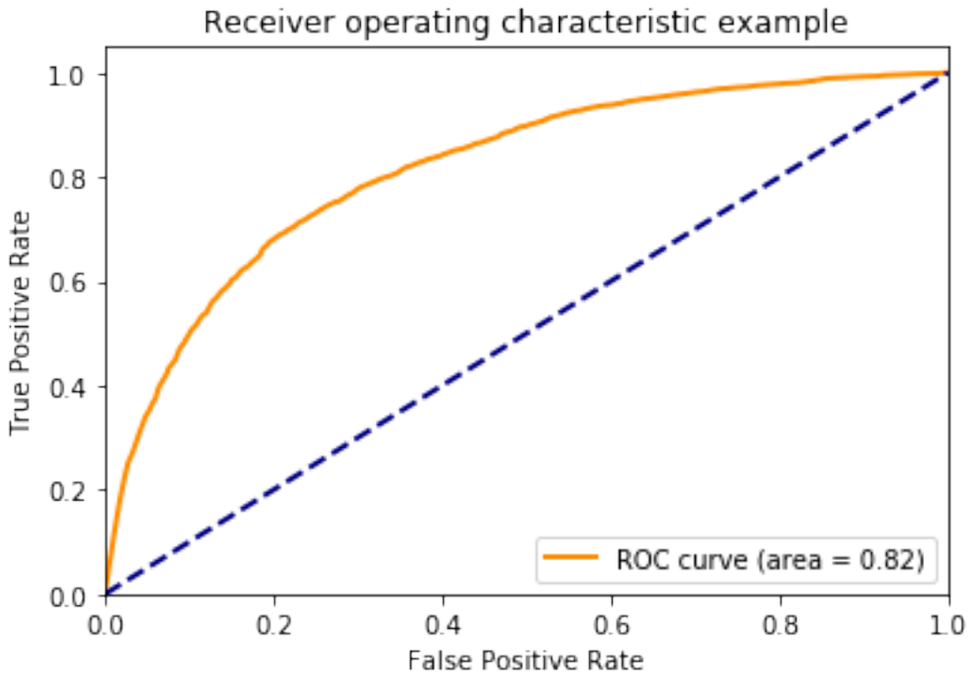
- The scatter matrix does not show any feature which will clearly separate out active vs. inactive users. Some features do seem to help.
- Thus, we probably have a highly non-linear relationship and should consider a **random forest** or **gradient boosted decision tree**, which are ensemble versions of a decision tree.

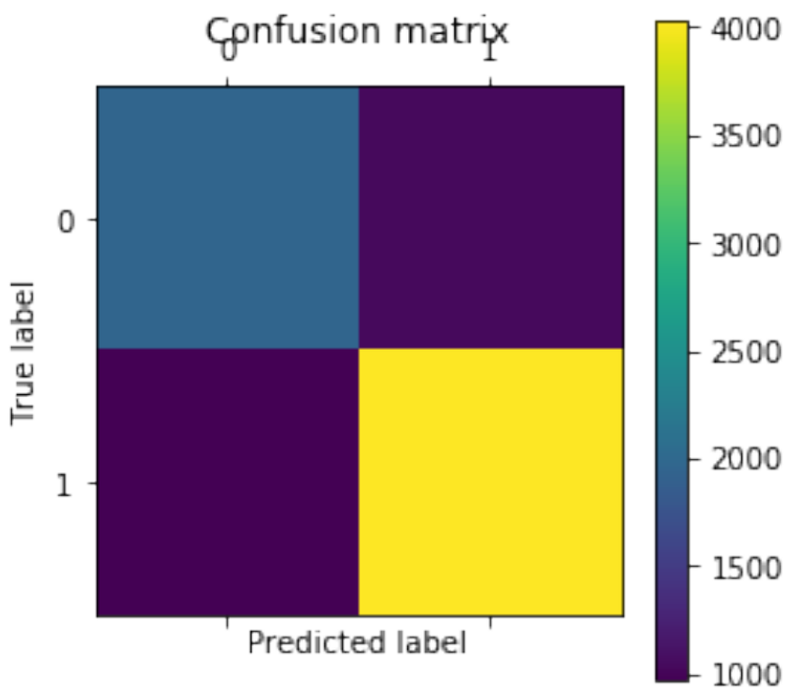
LOGISTIC REGRESSION



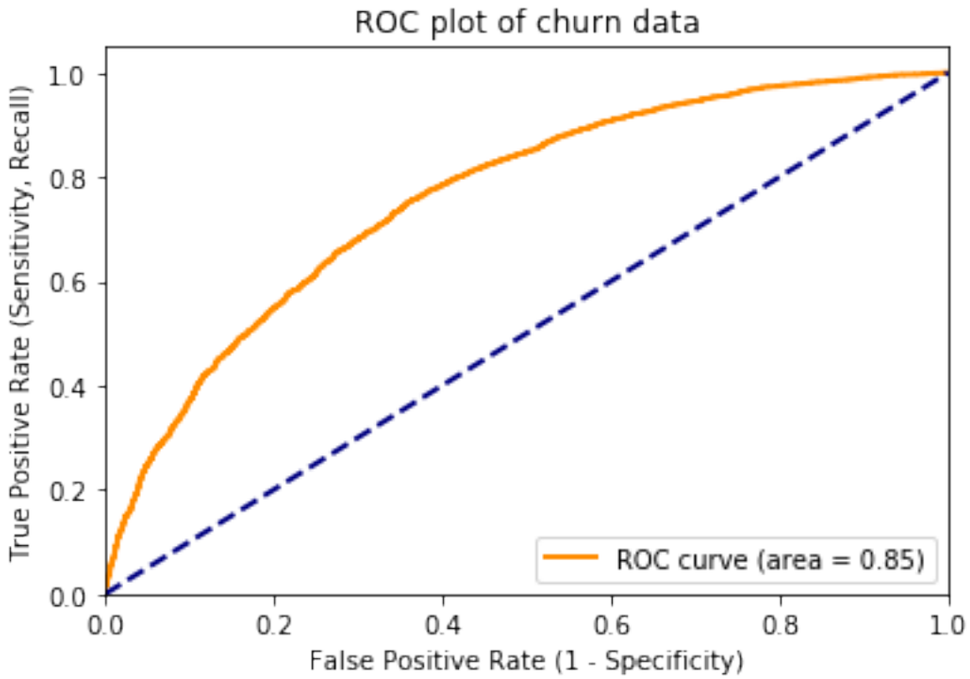


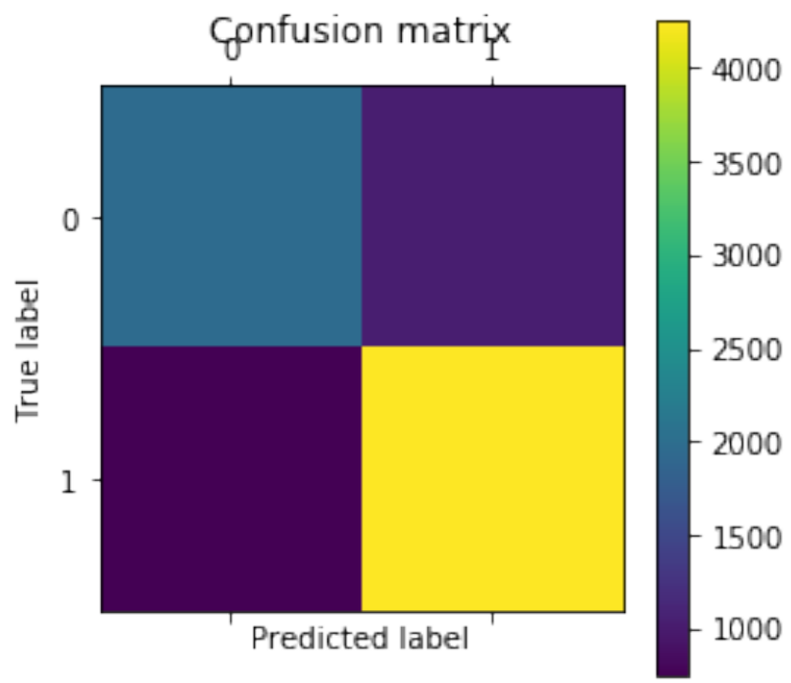
RANDOM FOREST





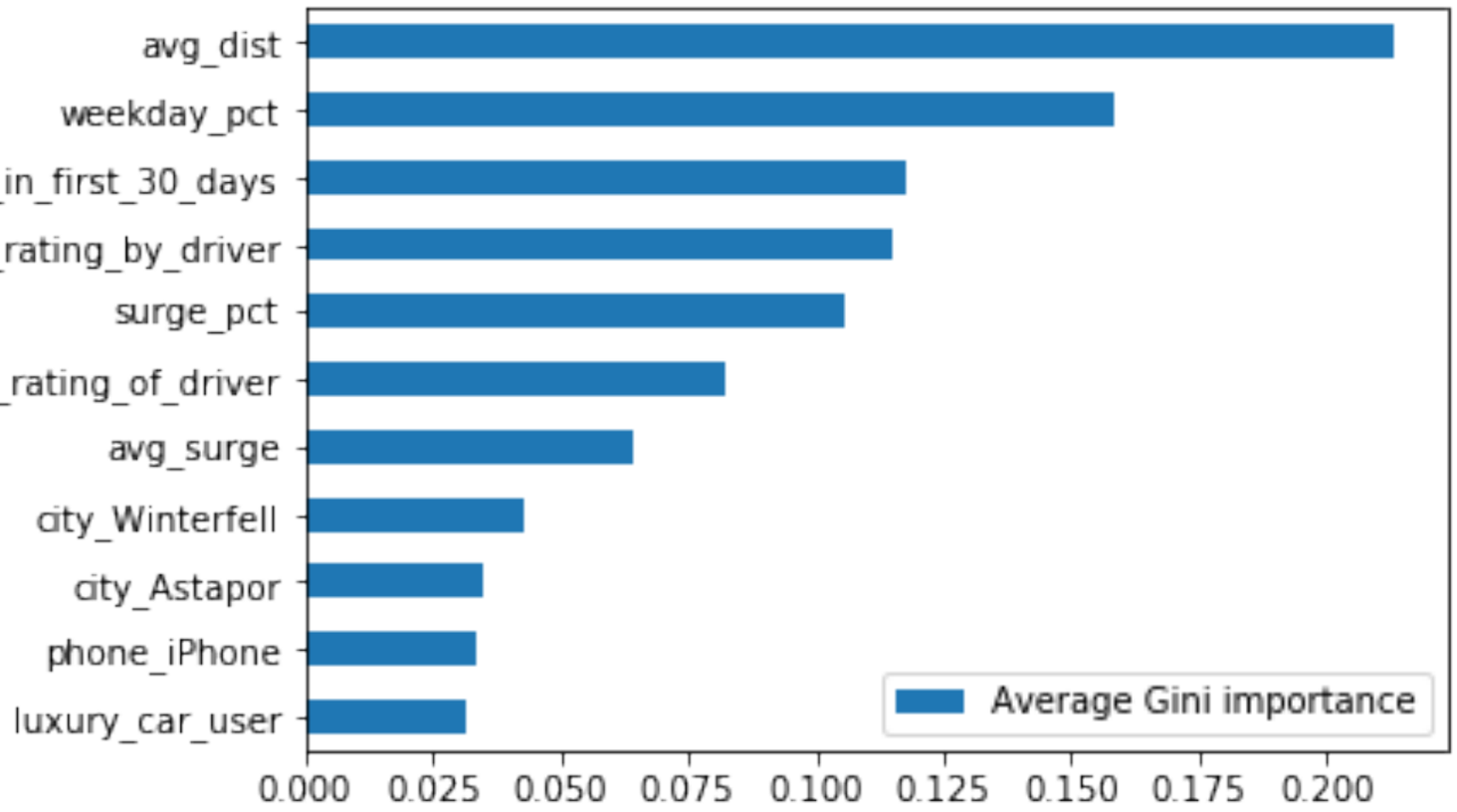
GRADIENT BOOSTING





This is the best and so it will be fine-tuned further via grid search

FEATURE IMPORTANCE



The three most importance features are:

avg_dist, weekday_pct, trips_in_first_30_days

COMPARISION

Model	Score	AUC
Logistic Regression	0.7161	0.7611
Random Forest	0.7494	0.8151
Gradient Boosting	0.7784	0.8485
Gradient Boosting (Grid Search)	0.7923	0.8548