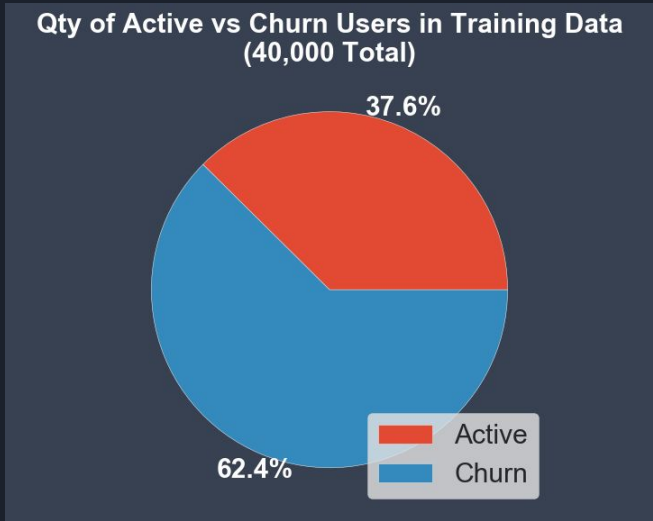
Reducing Churn with Supervised Learning

Min Tan
Jon Lee
Helen Tong
Temesgen Tesfay
Chelsea Ramos

Data Wrangling

- Target as 1
  - Trip date is above June 1st 2014
  - Meaning active user if last trip was on or after June 1
- Columns with null values:
  - Avg_rating_by_driver, Avg_rating_of_driver
    - Continuous data (1.0-5.0)
    - imputed NaNs with mean by city
  - Phone
    - Categorical data, used Pandas get_dummies
- Data Leakage
  - There were records after June 1st
  - Prediction data leak into training
  - CV using time series cross validation



Alternative Models on Training Data

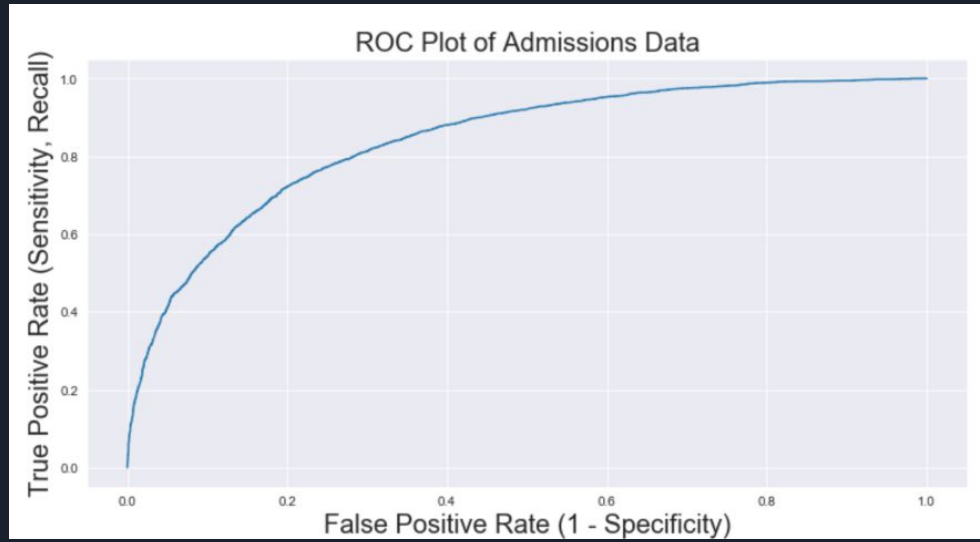
- Logistic Regression
  - Precision score: 0.67
  - Recall score: 0.48
  - Mean 5-fold cv score: 0.66
- Random Forest
  - Precision score: 0.73
  - Recall score: 0.66
  - Mean 5-fold cv score: 0.75
- Gradient Boost
  - Precision score: 0.74
  - Recall score: 0.66
  - Mean 5-fold cv score: 0.78

Final Model

- Gradient Boosting Classifier with Grid Search CV
  - Precision score: 0.706
  - Recall score: 0.672
  - Mean 5-fold cv score with training data: 0.78
- Features: Used all (all columns except target)
- Hyperparameters:
  - Loss: deviance
  - Max depth: 4
  - Max features: auto
  - Min leaf samples: 0.1
  - Estimators: 300

Performance Metrics

- Seek to maximize
- Average Precision Recall
  - .598
- ROC AUC
  - 0.75
- Accuracy not good
  - Imbalanced class

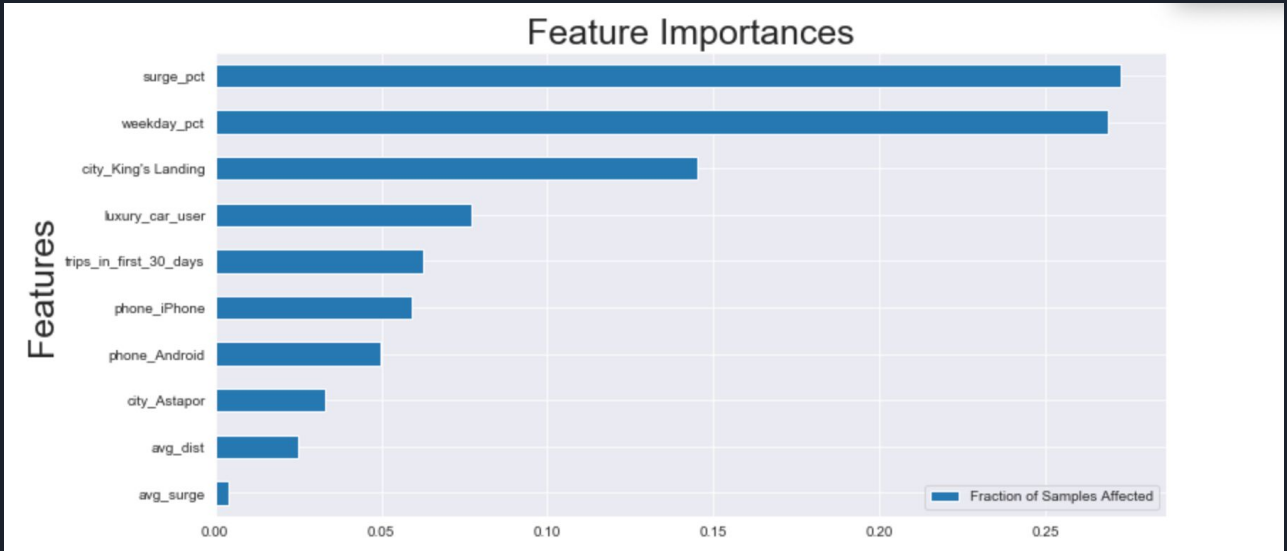


		Actual Class	
		Active	Churn
Predicted Class	Active	True Positives	False Positives
	Churn	False Negatives	True Negatives

		Actual Class	
		Active	Churn
Predicted Class	Active	2534	1238
	Churn	1054	5174

Plan Proposed to Reduce Churn

- Reduce surge times/pricing
- Promote more on weekdays/toward weekday commuters
- Promote more in King's Landing



Potential Impact

- Increase customer retention
  - Increase profit
  - Retain customer loyalty
- Reduce churn cost
- Predicted churn
  - Outreach campaign to retrieve churned customers

Planning

- How did you compute the target?
  a. Target: ‘active’ = 1 if ‘last_trip_date’ > June 1, 2014
- What model did you use in the end? Why?
  a.
- Alternative models you considered? Why are they not good enough?
  a. Logistic regression
  b. Decision tree
  c. Random forest
  d. Gradient boost/Adaboost
- What performance metric did you use to evaluate the *model*? Why?
  a. Precision, since data was imbalanced (~36.6 Active vs 63.4% Inactive), accuracy was not a good metric
- Based on insights from the model, what plans do you propose to reduce churn?
  a. Reduce surge times/pricing
  b. Promote more on weekdays/toward weekday commuters?
  c. Promote more in King’s Landing
- What are the potential impacts of implementing these plans or decisions? What performance metrics did you use to evaluate these *decisions*, why?
  a. Potential impacts of implementing plans or decisions