Case Study of Churn Prediction

Objective

- Predicting rider retention rate for a ride sharing company

- Propose solution to improve customer retention rate

Data Engineering

- Add a new column[‘Churn?’] based on two date columns, which were removed later to avoid data leakage

- Null values were replaced

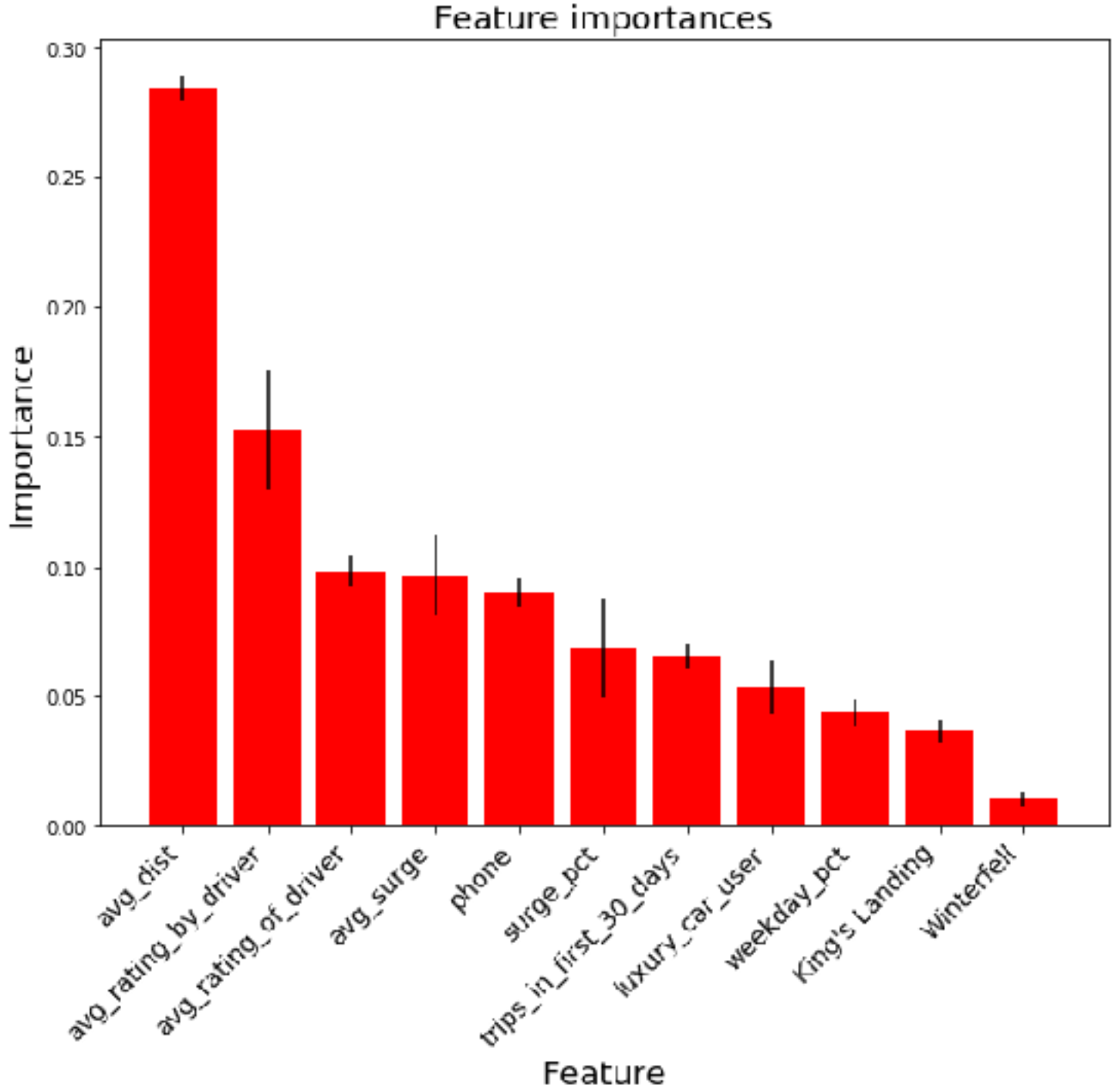
- Categorical features were converted to numeric type or replaced by several dummy features

Model - Random Forest

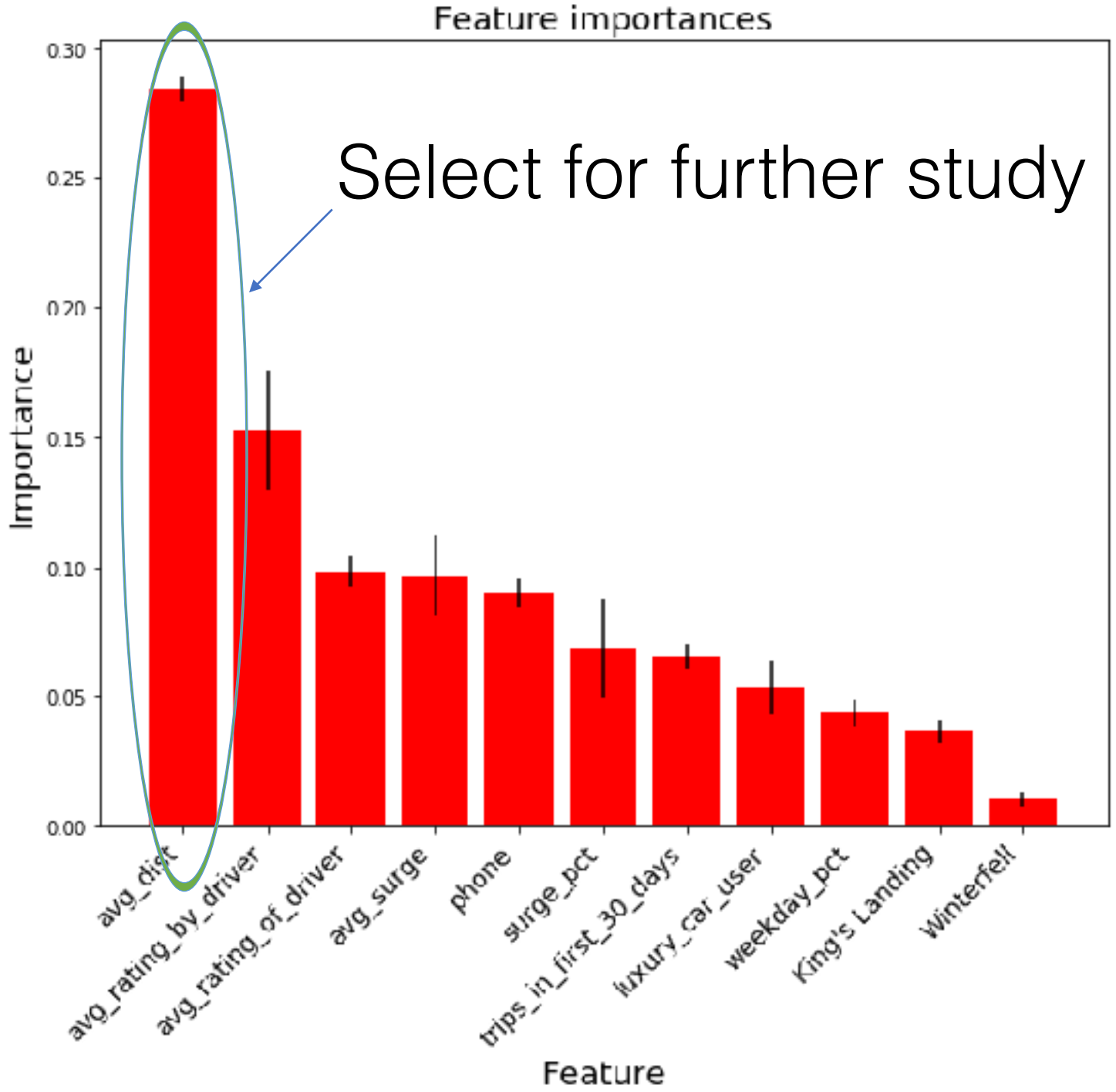
- Why?
  - Flexible
  - Easy explained
  - Fast implementation
  - Works well for non-linear classification

- How?
  - Study feature importances
  - Find the optimal parameters by cross-validated grid search
  - Fit optimal model
  - Test model with confusion matrix and ROC curve

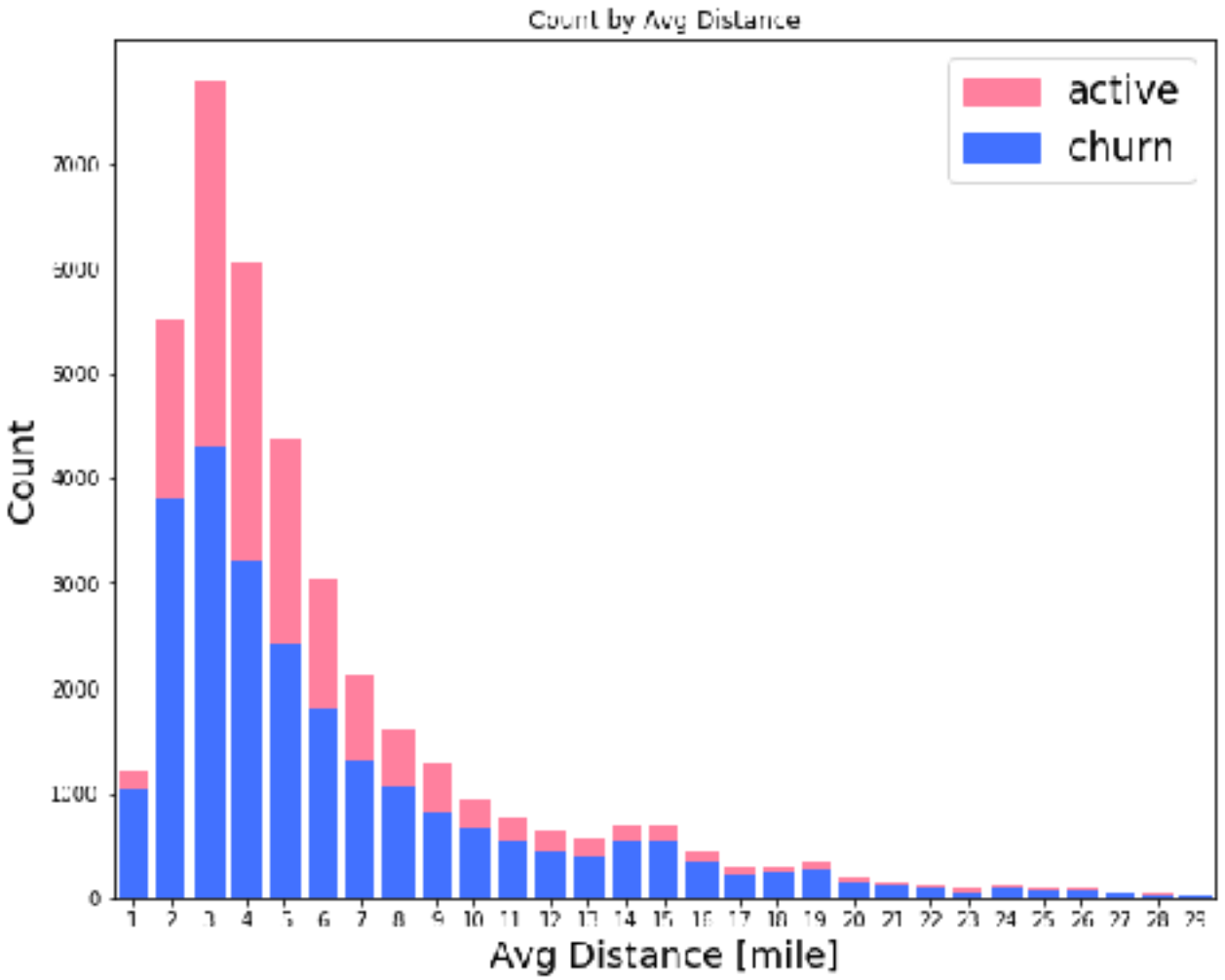
Feature Importances via Random Forest

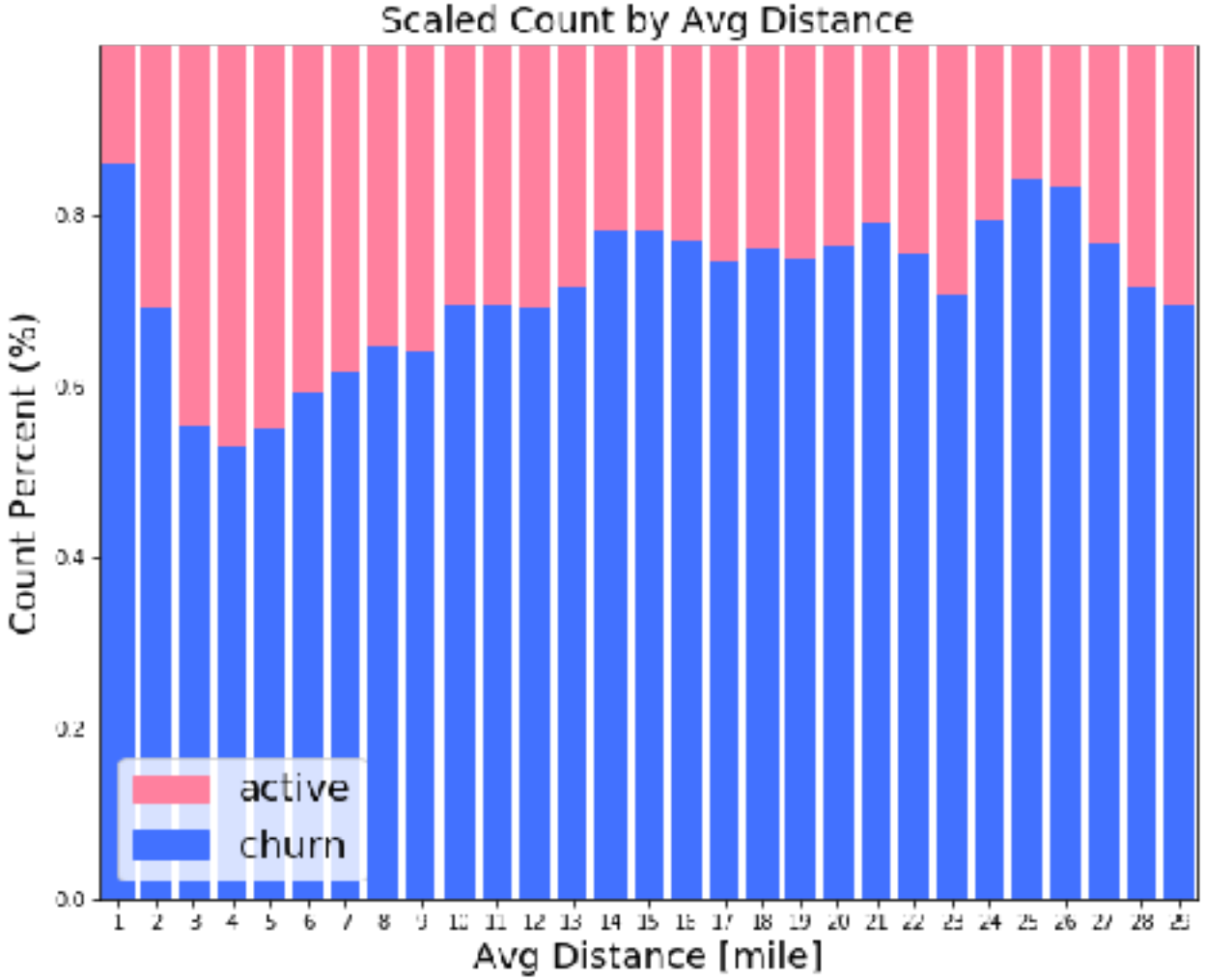


Feature Importances via Random Forest

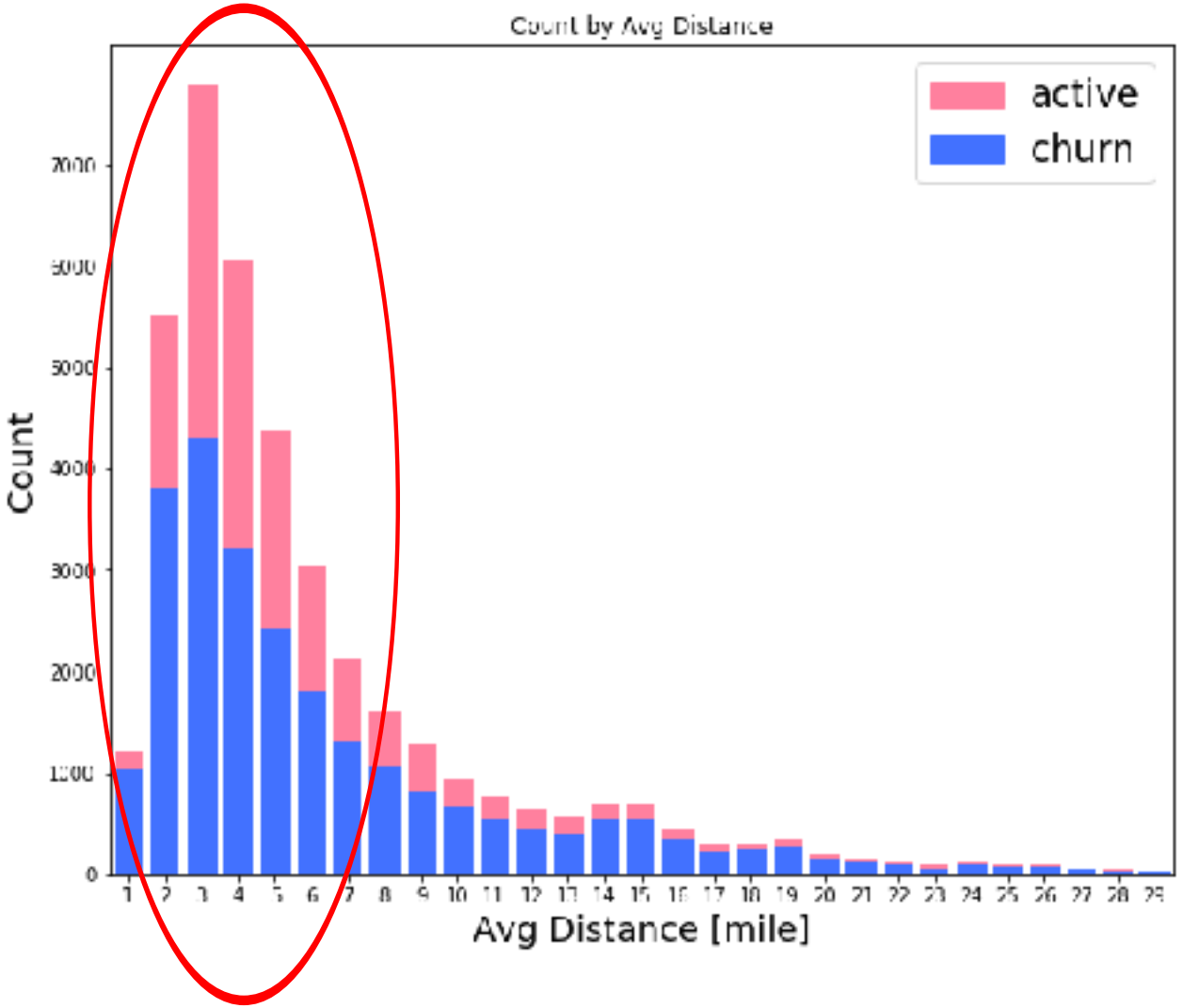


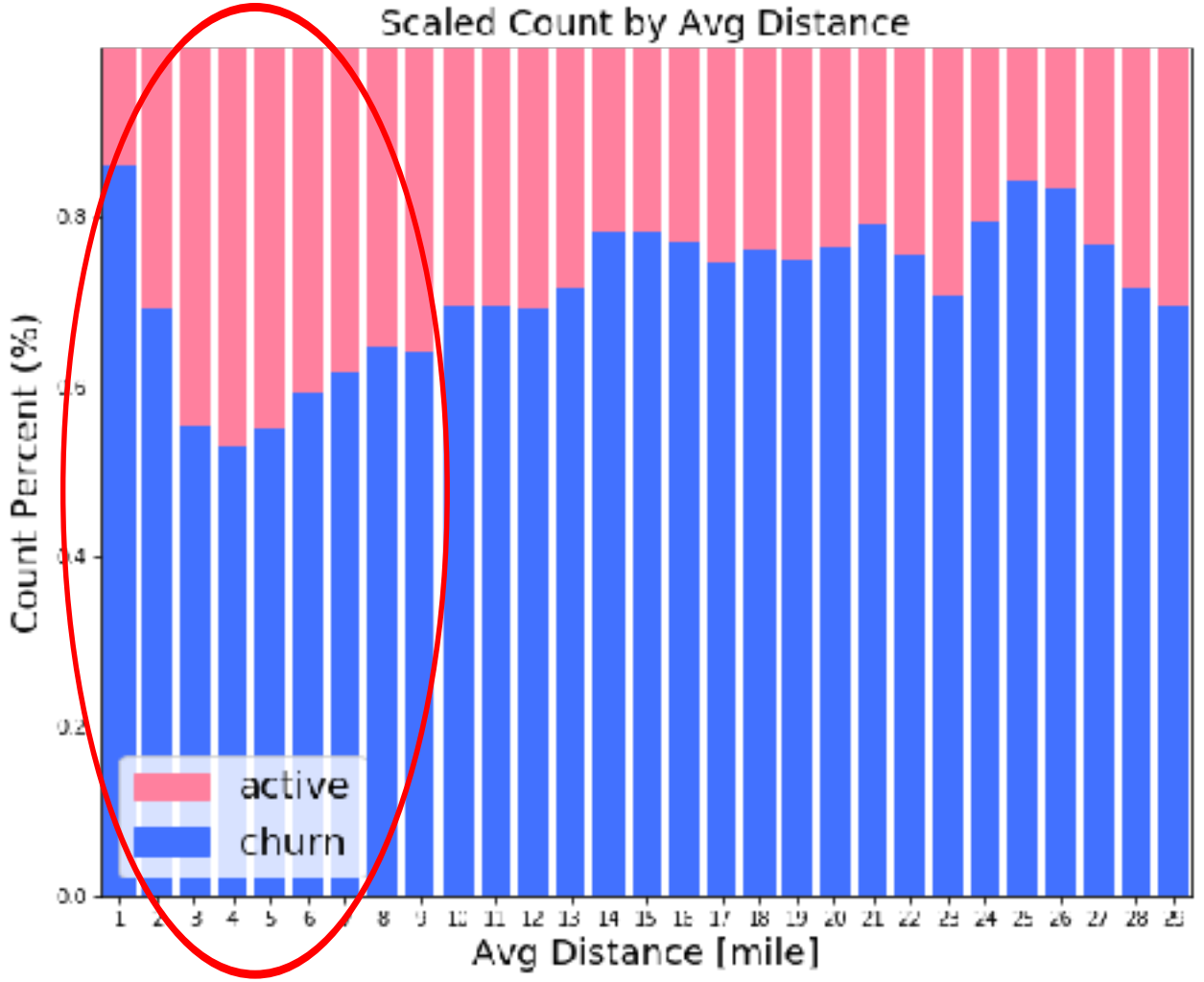
Average distance and churn relationship



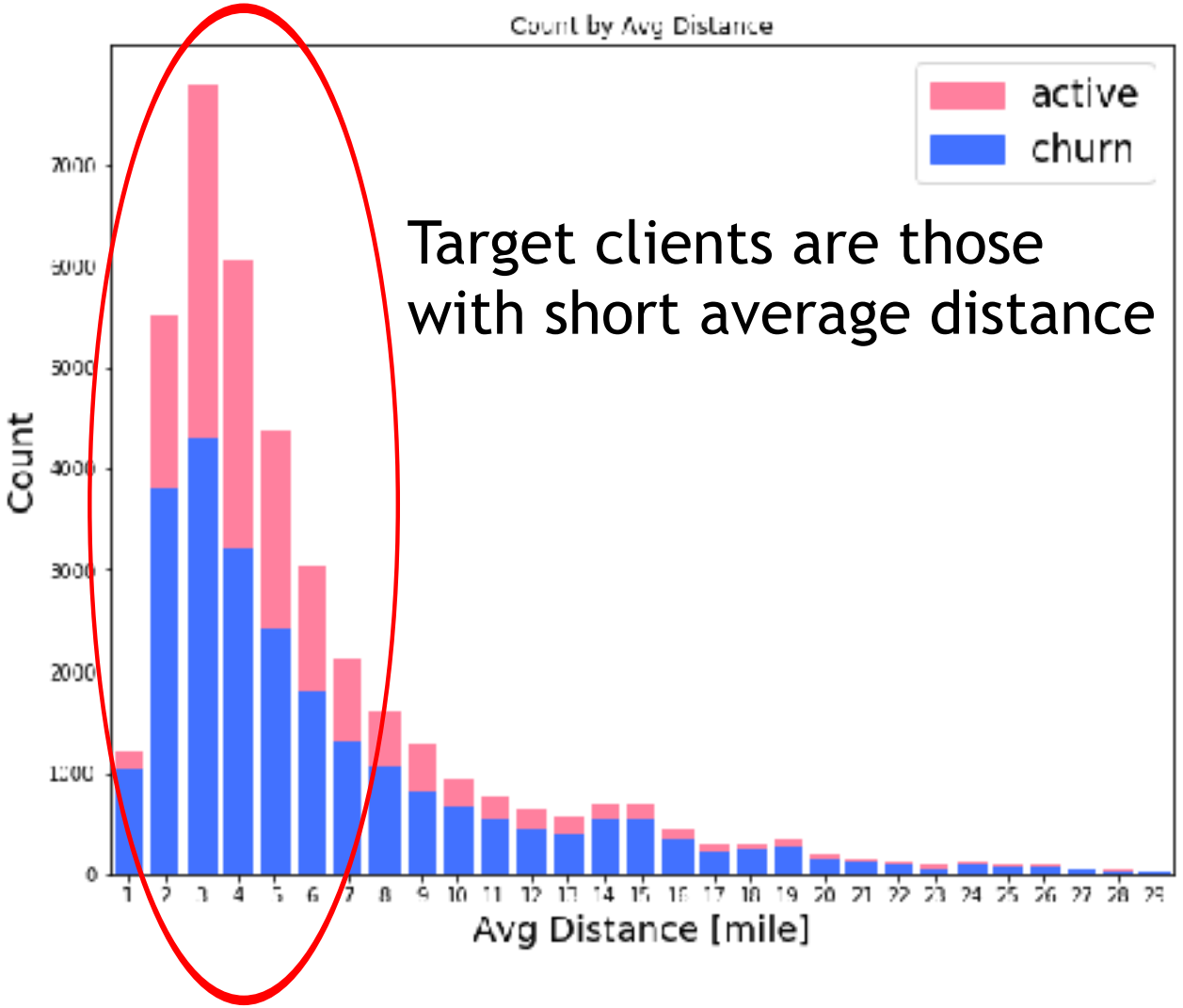


Average distance and churn relationship

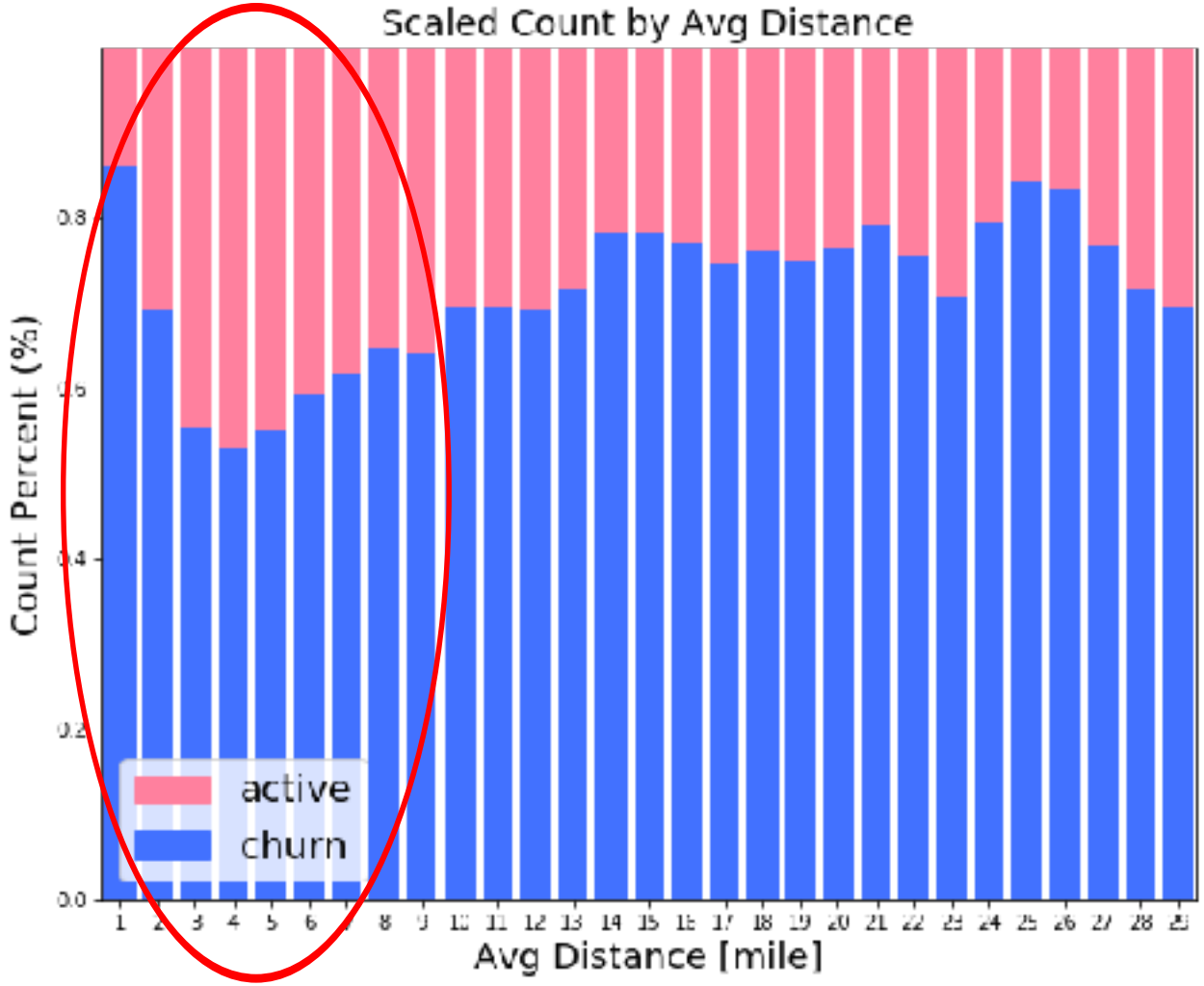




Average distance and churn relationship

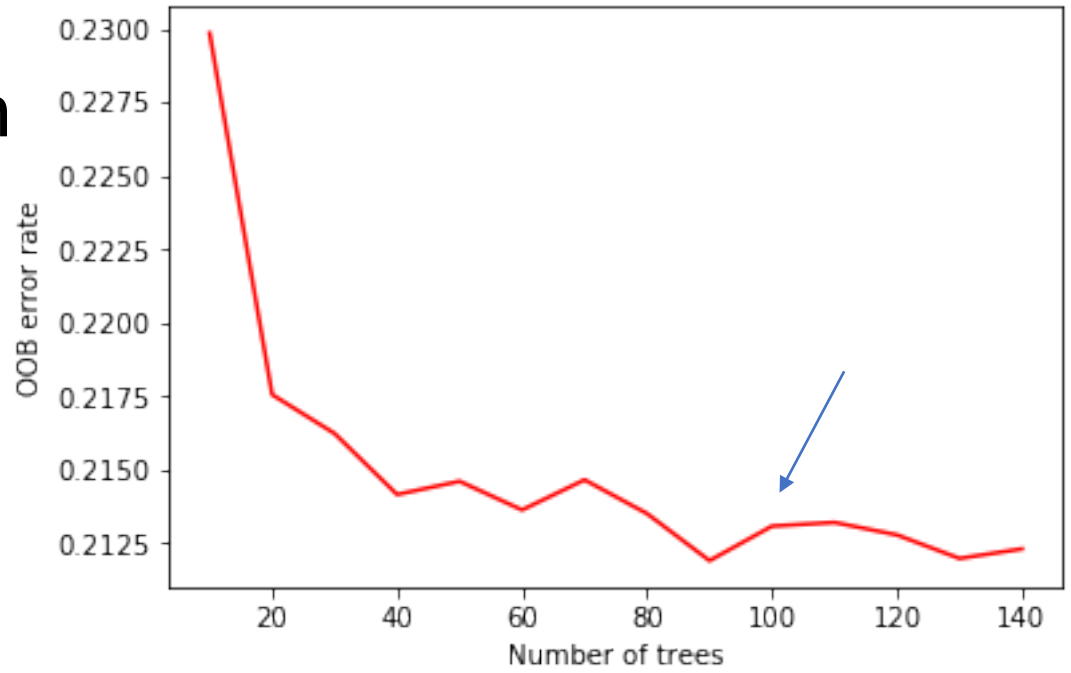


Target clients are those with short average distance



Model

- Get the optimal parameters by searching over gird parameter space
- Determine the number of trees based on plot of OOB error vs tree number
- Optimal parameters
  - number of tree = 100
  - max_features = 3
  - min_samples_leaf = 14

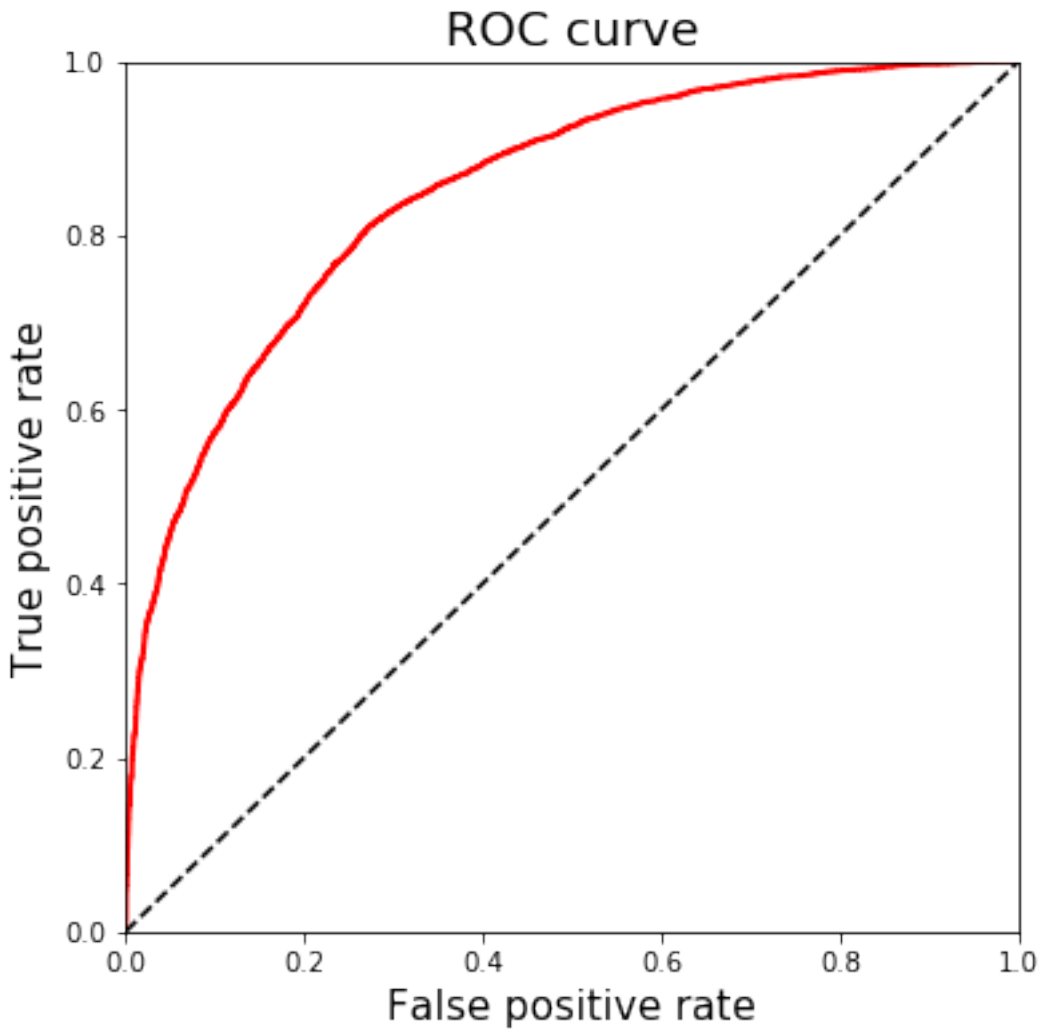


Metrics

Accuracy score on test:	0.778
Out of bag score:	0.787
Precision on test:	0.795
Recall on test:	0.866

Confusion Matrix on Test data

	Predict Churn	Predict Active
Act Churn	**5392**	**836**
Act Active	**1389**	**2383**



Benefit Matrix Simulations

Benefit Matrix

	Predict Churn	Predict Active
Act Churn	**$36**	**$0**
Act Active	**$180**	**$200**

Confusion Matrix on Test data

	Predict Churn	Predict Active
Act Churn	**5392**	**836**
Act Active	**1389**	**2383**

ASSUMPTIONS

1. Revenue per customer $200/ month
2. 10% discount for those who we predicted Churn
3. With 10% discount, 20% of clients will stay

Benefit Matrix Simulations

Benefit Matrix

	Predict Churn	Predict Active
Act Churn	**$36**	**$0**
Act Active	**$180**	**$200**

Confusion Matrix on Test data

	Predict Churn	Predict Active
Act Churn	**5392**	**836**
Act Active	**1389**	**2383**

ASSUMPTIONS

1. Revenue per customer $200/ month
2. 10% discount for those who we predicted Churn
3. With 10% discount, 20% of clients will stay

CONCLUSION

Prefer Type I error to Type II error

Conclusion

- Build a model of churn prediction using Random Forest
- Study the feature importances
- Target clients with short average distance (< 8 miles)
- The model was evaluated using some metrics, e.g. confusion matrix, ROC curve, etc
- A benefit matrix was studied

Future work

- Try other models
- Study other important features
- Collect more data regarding trips made by each user
- Recommend creating a new campaign, for example,
  - Offer discount to users who travel long distance
  - Provide membership reward service