CAPSTONE PROJECT: PREDICTING CHURN CUSTOMERS

Albert Chiu
December 2018

PROBLEM

- Retaining existing customers is important for subscription businesses.
- Existing customers:
  - reduce expenditure on marketing
  - provide free word of mouth advertising
  - more likely to pay for premium features and products

Churn is the annual rate at which customers stop subscribing to a service.

DATA SET

- Data set: *https://www.kaggle.com/blastchar/telco-customer-churn*

- The data set used for this project contains customer information for a telecom company, including **7043 customer observations** with **21 features**

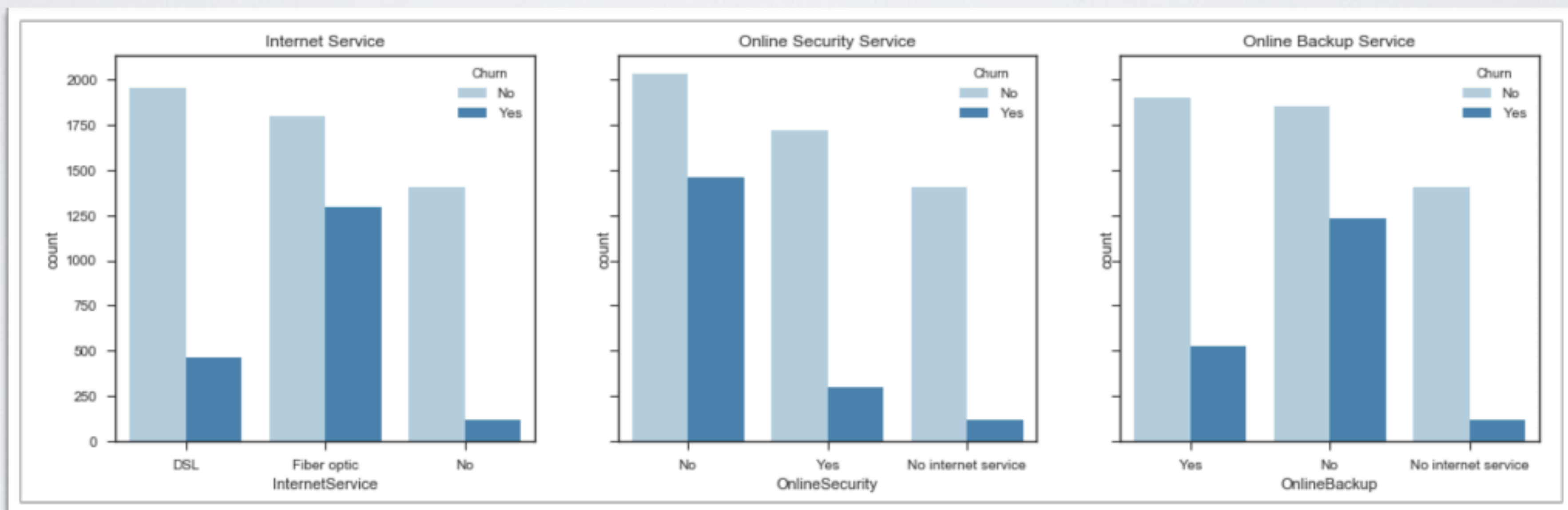
- As expected, there is a data set imbalance: **74% non-churn** customers compared to 26% **churn customers**

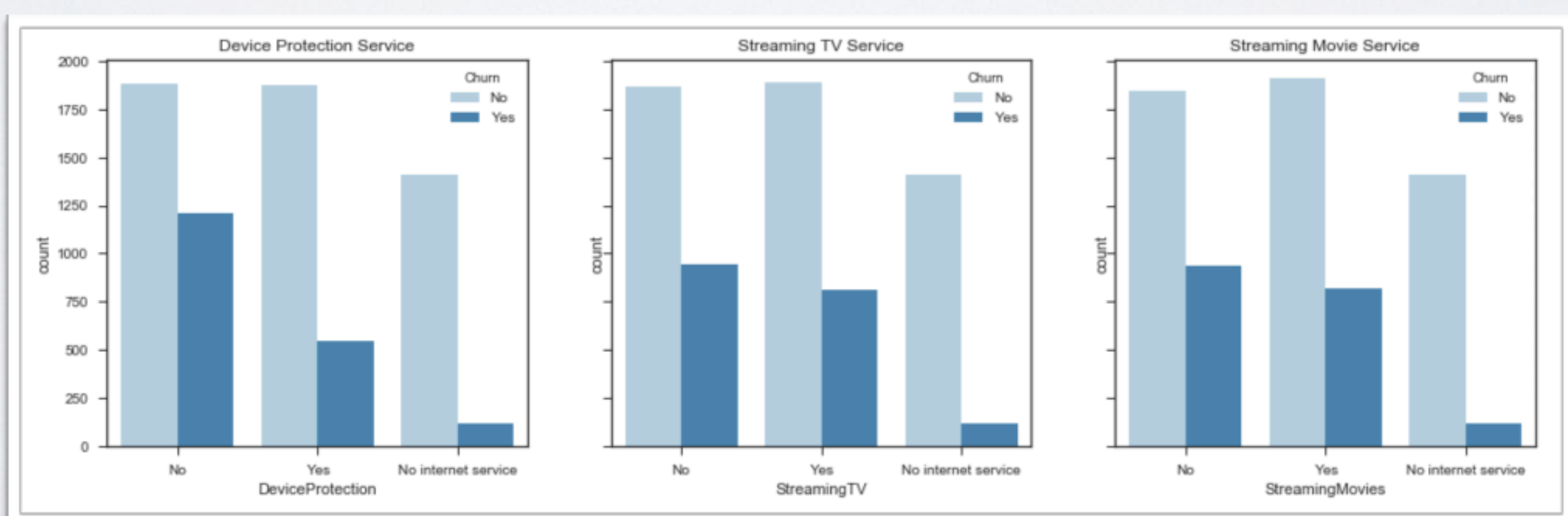
PROJECT GOALS

1. Identify the profile of customers that are likely to churn
2. Build models that predict probability of whether customers will churn or not

CHURN CUSTOMERS

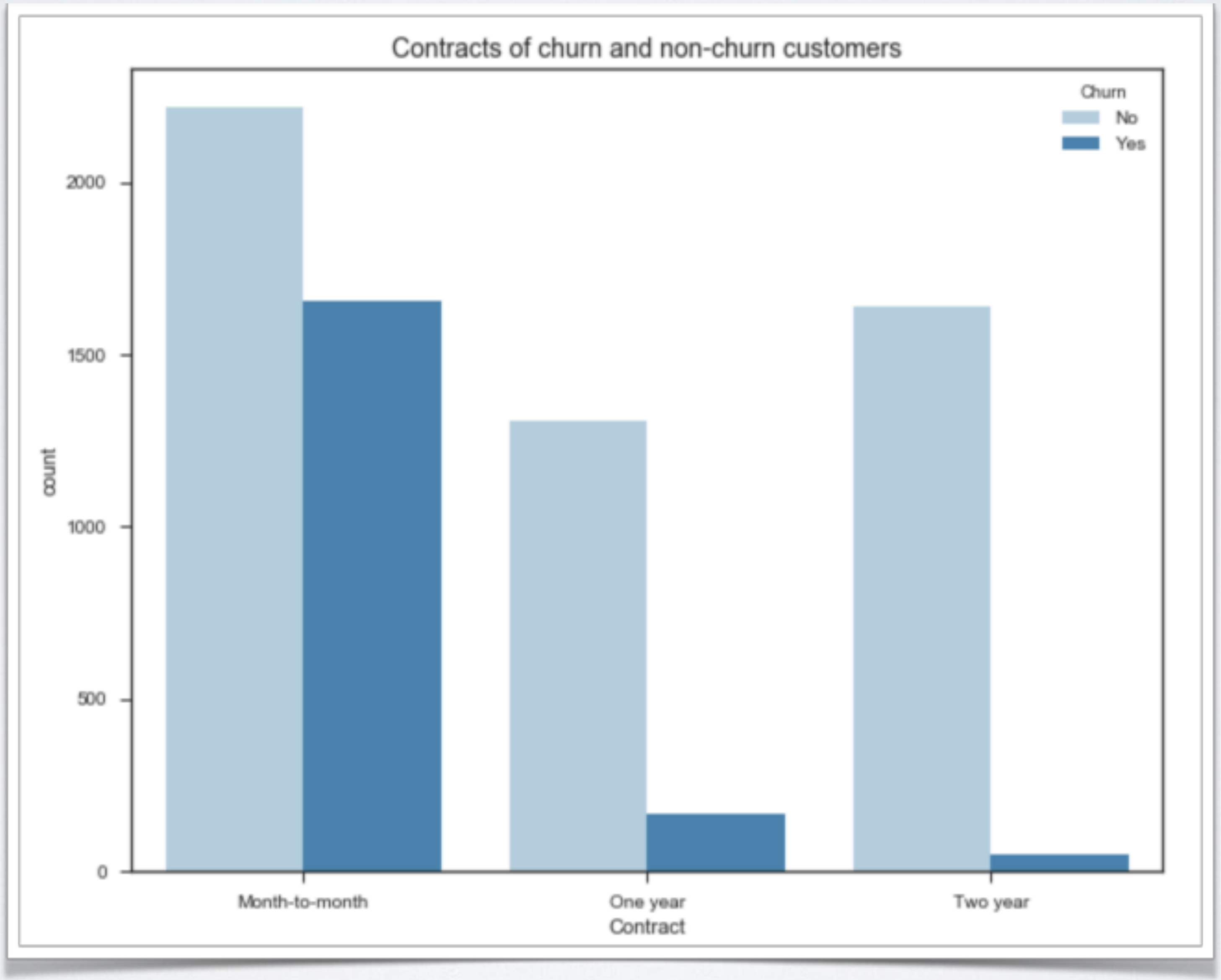
- It was observed that churn customers primarily subscribe to fiber optic internet with few other services besides streaming TV and movies





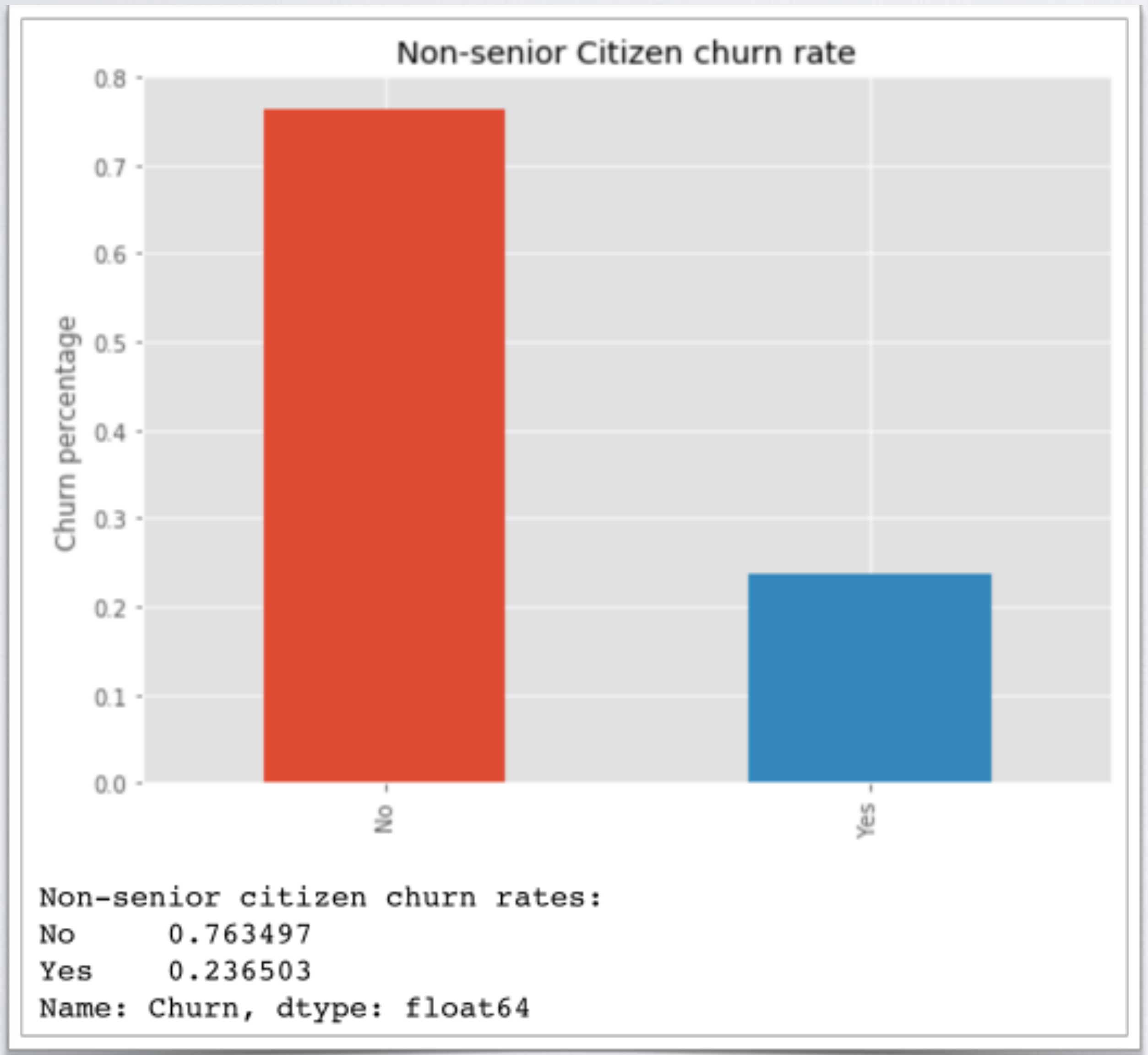
CHURN CUSTOMERS

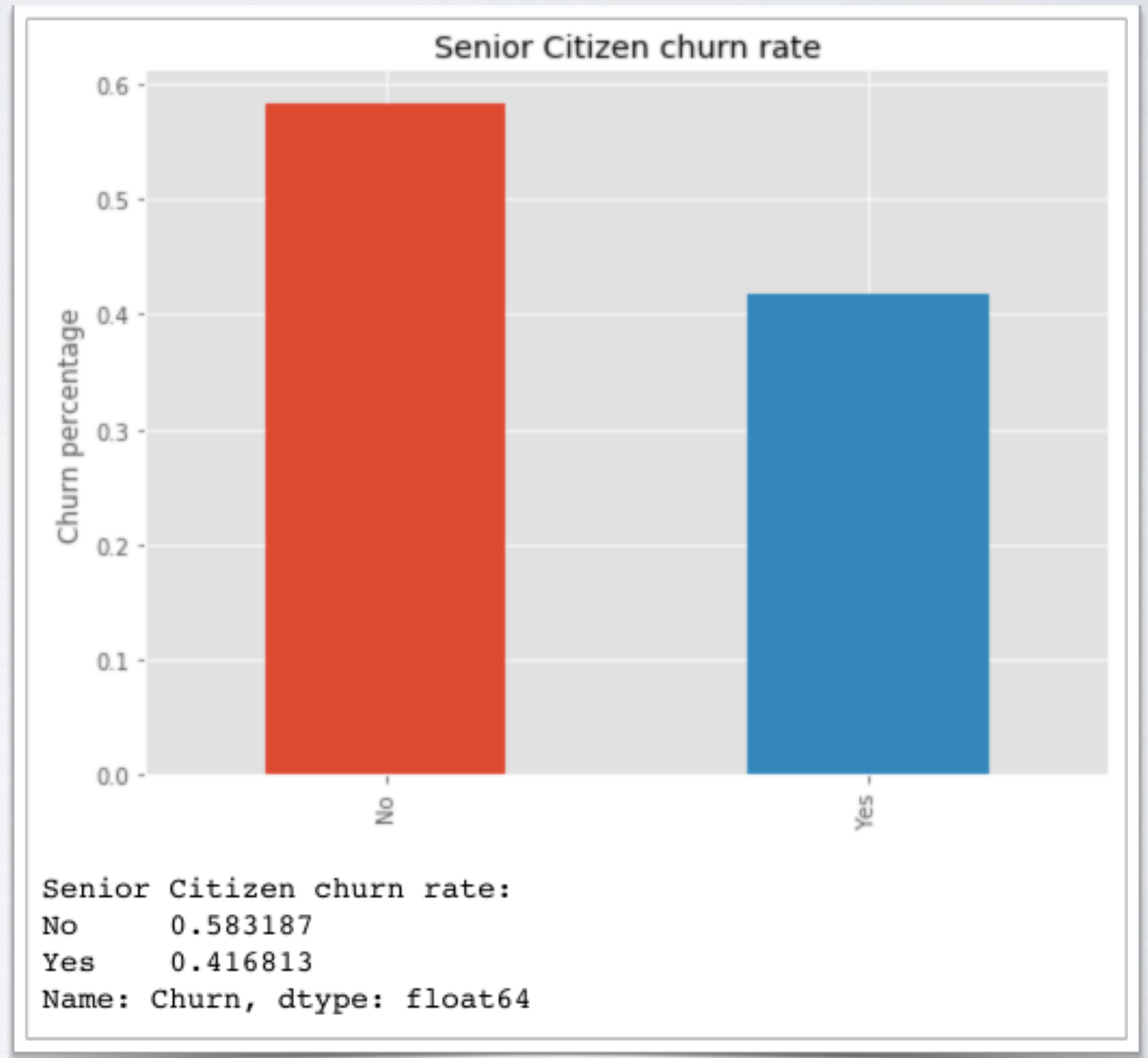
- Churn customers overwhelmingly favor month-to-month contract.
- Month-to-month contracts account for 88% of the contract type for churn customers compared to 43% for non-churn customers.



CHURN CUSTOMERS

- Although senior citizens account for around 20% of the customer base, they churn nearly twice as often as non-senior citizens.





MACHINE LEARNING MODELING

- Since the machine learning models for this model will be making categorical predictions (whether customers churn or not), **supervised classification methods** were used.

- **Churn recall** will be used to measure algorithm performance, since the company will be interested in identifying as many potential churn customers as possible.

MACHINE LEARNING MODELING

- Hyperparameter tuning was used to optimize results.
- To address data set imbalance, resampling techniques were applied. This improved model performance significantly.

MODEL CLASSIFICATION REPORTS

Classification report results for non-churn customers:

	Precision	Recall	F-score	Support
k-NN	0.7872	0.9404	0.857	1007
Lasso Reg	0.8249	0.9027	0.862	1007
Ridge Reg	0.8223	0.9007	0.8597	1007
Random Forest	0.8116	0.9067	0.8565	1007
Random Forest w/ Random Over-Sampler	0.8204	0.8709	0.8449	1007
Random Forest w/ SMOTE	0.7982	0.8918	0.8424	1007
Log Reg w/ Random Over-Sampler	0.912	0.7408	0.8175	1007
Log Reg w/ SMOTE	0.9096	0.7398	0.816	1007
Random Forest w/ Random Under-Sampler	0.8708	0.7498	0.8058	1007
Random Forest w/ Tomek Links	0.8313	0.8858	0.8577	1007
Log Reg w/ Random Under-Sampler	0.9101	0.7239	0.8064	1007
Log Reg w/ Tomek Links	0.8427	0.862	0.8522	1007

Classification report results for churn customers:

	Precision	Recall	F-score	Support
k-NN	0.7059	0.36	0.4768	400
Lasso Reg	0.6787	0.5175	0.5872	400
Ridge Reg	0.6711	0.51	0.5795	400
Random Forest	0.6667	0.47	0.5513	400
Random Forest w/ Random Over-Sampler	0.6154	0.52	0.5637	400
Random Forest w/ SMOTE	0.6135	0.4325	0.5073	400
Log Reg w/ Random Over-Sampler	0.5569	0.82	0.6633	400
Log Reg w/ SMOTE	0.5544	0.815	0.6599	400
Random Forest w/ Random Under-Sampler	0.5333	0.72	0.6128	400
Random Forest w/ Tomek Links	0.6557	0.5475	0.5967	400
Log Reg w/ Random Under-Sampler	0.5413	0.82	0.6521	400
Log Reg w/ Tomek Links	0.6313	0.595	0.6126	400

MACHINE LEARNING RESULTS

- The top performing classifier was Logistic Regression paired with SMOTE, random oversampling or random random undersampling.

- There appears to be an inverse correlation between churn recall and churn precision.
  - A lower churn precision is an acceptable trade off for better churn recall performance. A company would prefer to identify as many potential churn customers as possible at the expense of slightly more incorrect churn predictions.

RECOMMEDATIONS

- Given the insights gained, the company can take some steps to help reduce churn rate:

  1. Conduct interviews and surveys to identify promotions preferred by senior citizens.
  2. Offer price promotion for service extension to high risk churn customers prior to a year and a half of tenure (average churn customer stops service at 18 months).
  3. Collect age of customers. Different age groups have different consumption habits. This information will allow for more granular insights of churn customers by age range.

SUPPLEMENTAL INFO

- For all my code and final report, please see my Github link for this project: *https://github.com/albertdchiu11/Springboard-Data-Science/tree/master/Capstone-1-Project*