How can telecom companies retain their customers?

An experiment with customer churn prediction

by

LaShawn Gaines

Why is a customer churn prediction model needed?

- Mobile telecommunications industry is highly competitive because the consumers have the power
- Estimates show that 97% of all Americans own a smartphone
  - Adding new customers requires attracting them from other service providers
- Maintaining market share must also therefore include a customer retention / churn reduction strategy

About the dataset

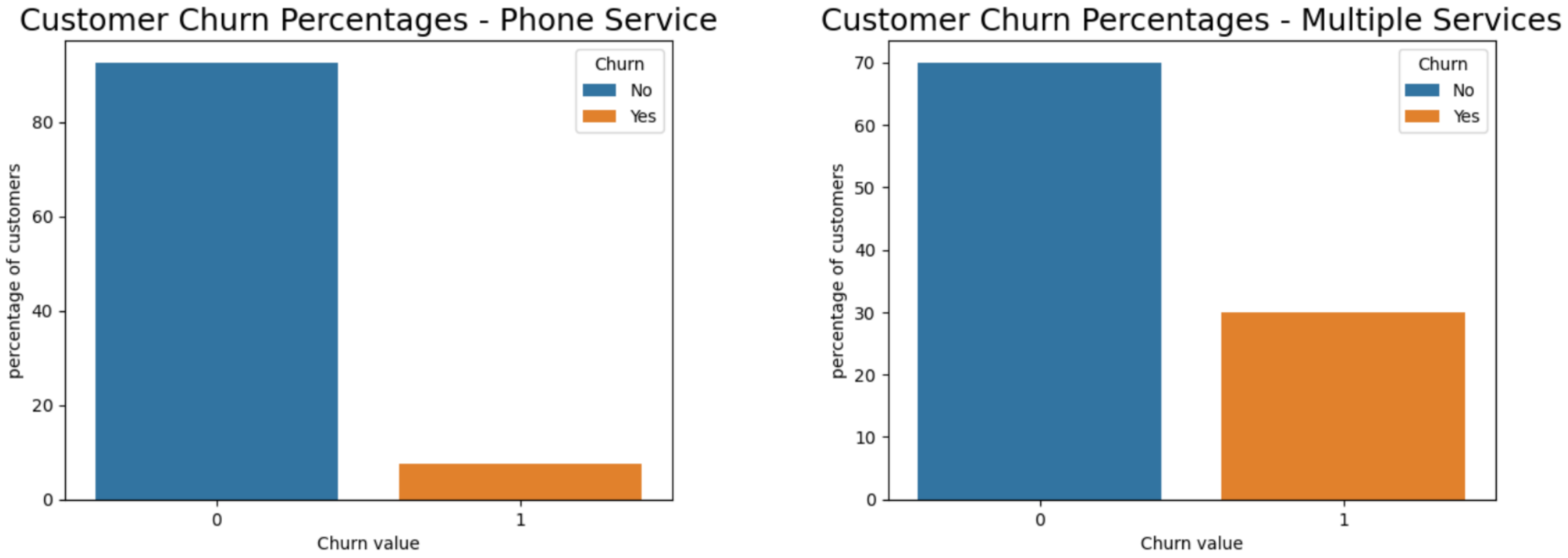
- IBM sample dataset located on kaggle
- Usability score of 8.82
- Contains sample of 7,043 customer records and 21 columns
    - customers that left company within the last month
    - services in which the customers subscribed (phone, internet, streaming)
    - customer account information (tenure, contract type, pay methods, monthly charges)
    - customer demographic data (age, gender, single vs partnered, dependents)

Data Wrangling

- High Kaggle usability score minimized the amount of cleaning
- Total charges was converted from object dtype to float
- IsSeniorCitizen was converted to object from numeric to match other categorical fields
- Converted service categorical data to numeric to create new bundled services categories
    - bundled phone & internet
    - count of total services each customer was subscribed
- Dropped 11 rows missing total charge information
- Dropped the customer ID

EDA – preliminary churn findings

- churn rate across all customers is 26.58%
- Phone services appear to have a higher retention rate than customers with multiple services

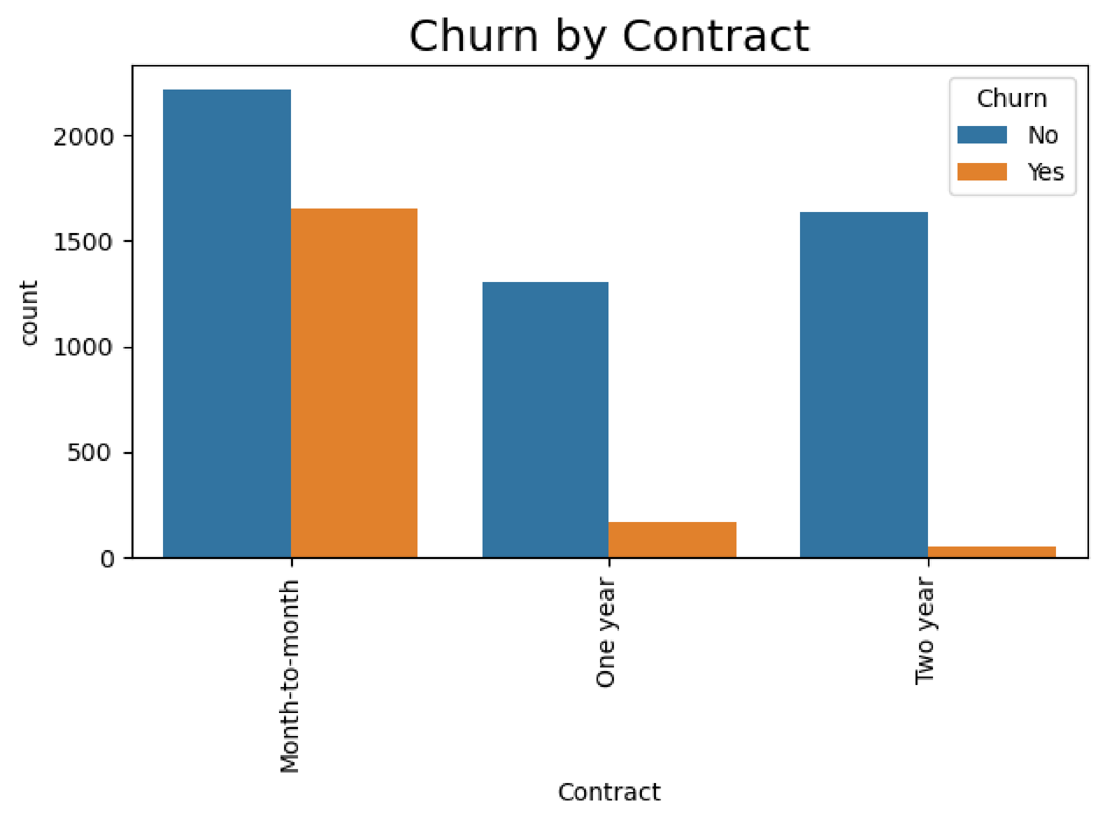


Churn Percentages - Phone Services
No: 92.57%
Yes: 7.43%

Churn Percentages - Multiple Services
No: 70.01%
Yes: 29.99%

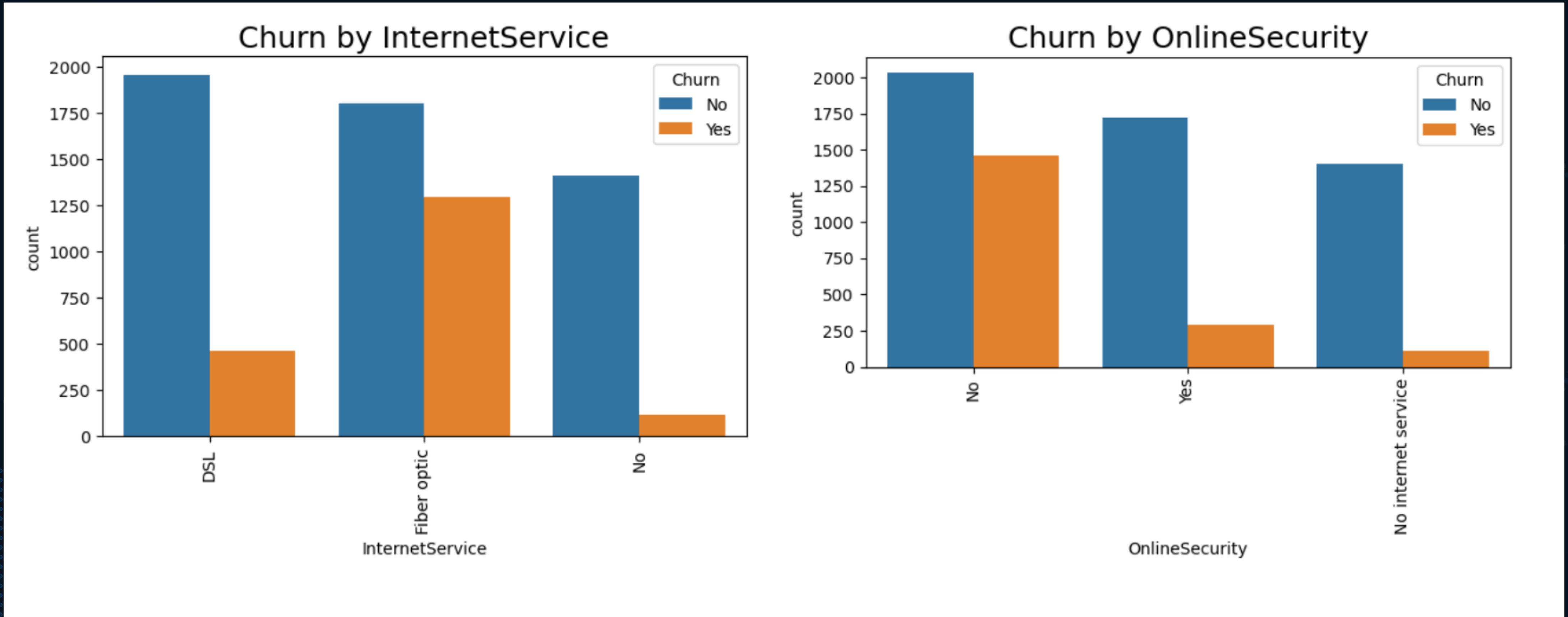
EDA – bivariate analysis

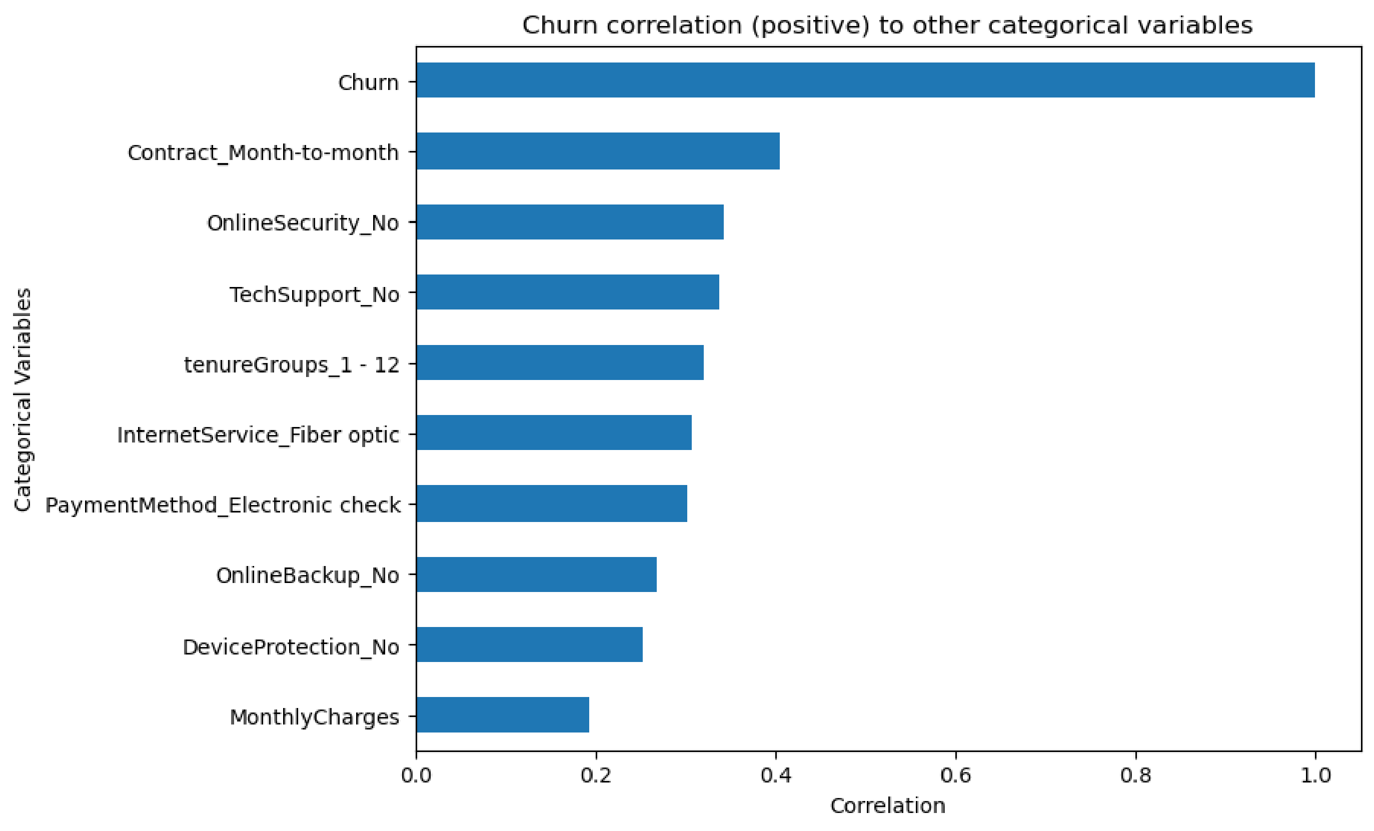
- Customers without long-term contracts are churning at a higher rate (lower barriers to disconnect)
- Month-to-month contracts are now the norm in the industry
- Further support for churn prediction



EDA – bivariate analysis of service offerings

- Fiber optic internet customer are more likely to churn (potential network performance issues)
- Online security customers are less likely to churn (opportunities for future bundled service offerings)



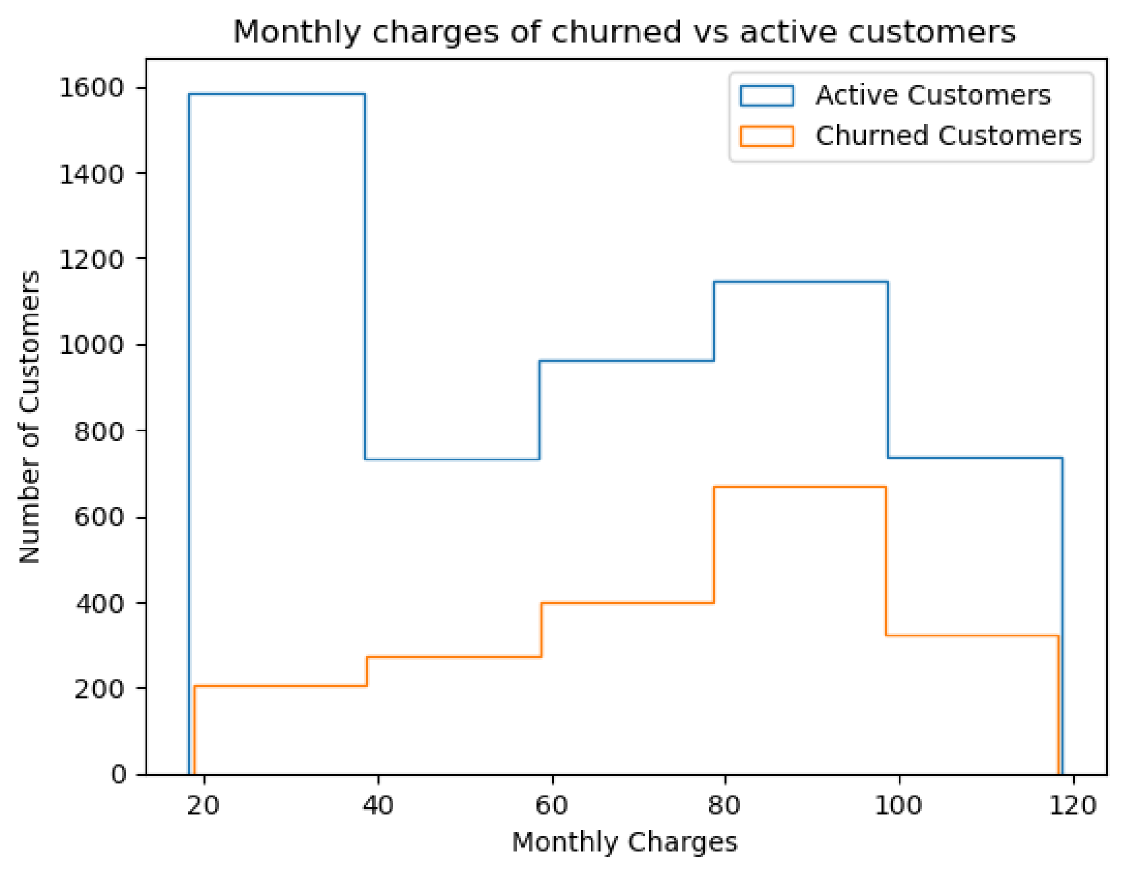


EDA – correlation analysis

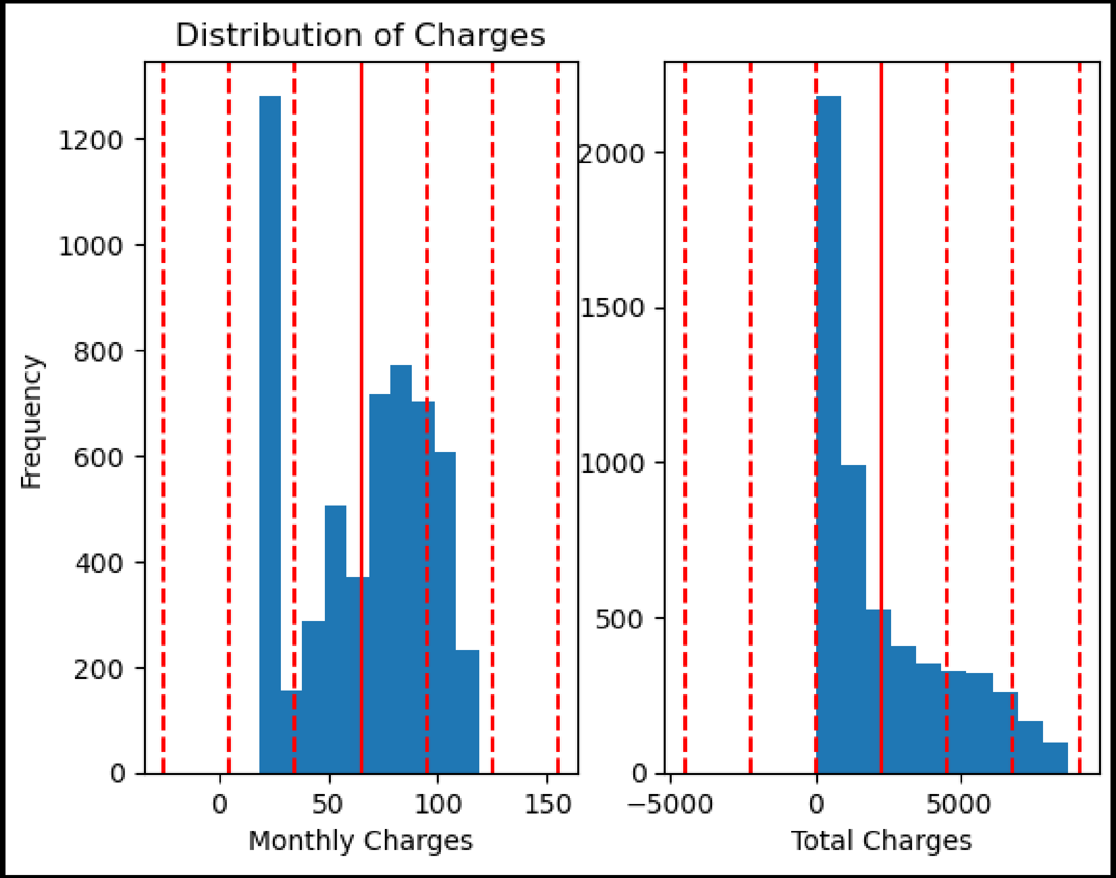
- strongest correlation is monthly contracts
- weakest correlation with monthly charges
- Not subscribing to ancillary services (online security and tech support) are moderately correlated

EDA – monthly charges vs customer status

- Churned customers with monthly charges between $80 - $100 have highest frequency
- Active customers are more frequently paying between $20 - $40 per month
- Business leaders may want to look at price/mix strategies to maximize revenues



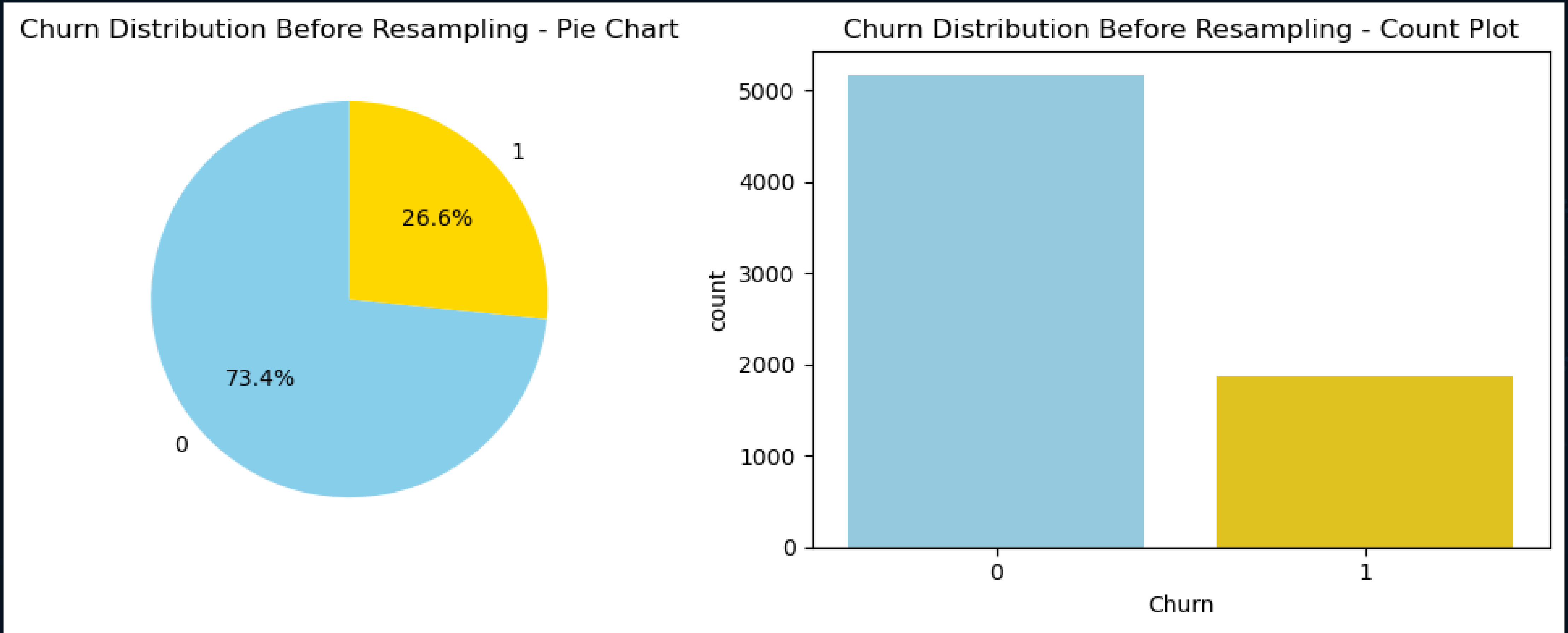
Preprocessing Steps



- Dummy encoding (dropping the first feature) performed for all categorical fields, other than churn
- Split data into training and test sets
- Scaled the data using sklearn’s StandardScaler
  - Monthly and Total charges are bi-modal and right skewed, respectively
  - All categorical fields are binary, thus posing a magnitude issue

Preprocessing Steps: resampling

- Given the nature of churn, data set was highly imbalanced
- Given the relatively small dataset we chose an over-sampling technique
  - SMOTE + Tomek Links
  - Resulting in a 50% majority to 50% minority

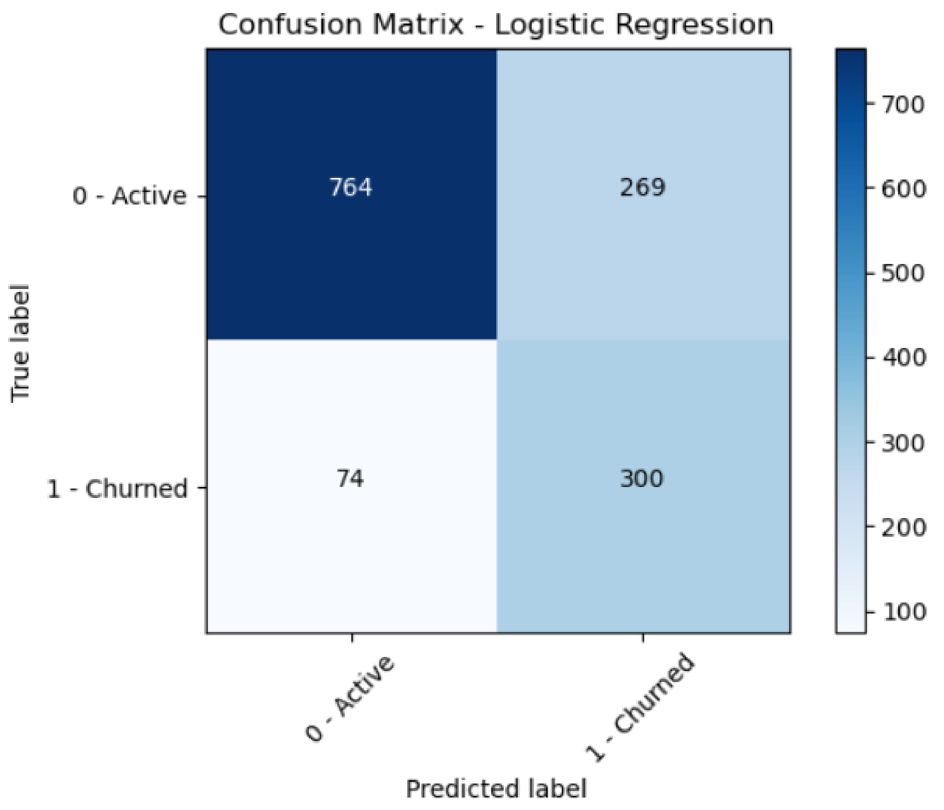


Modeling: churn prediction

- Algorithms used (classifiers)
  - Logistic Regression
  - Random Forest
  - Gradient Boosting
  - K-nearest Neighbors
  - Support Vector Classifier
- Established a baseline running each algorithm with default params
- Used GridSearch with 5-fold cross validation for hyperparameter tuning
- Scored models based on recall, choosing to minimize the number of false negatives

Modeling: winning algorithm is Logistic Regression

- Recall scores for churn were best with this model
- Minimizes the number of false negatives
- Trade-off is the number of false positives is quite high
    - potential for expending a lot of human capital to contact customers who are not about to churn

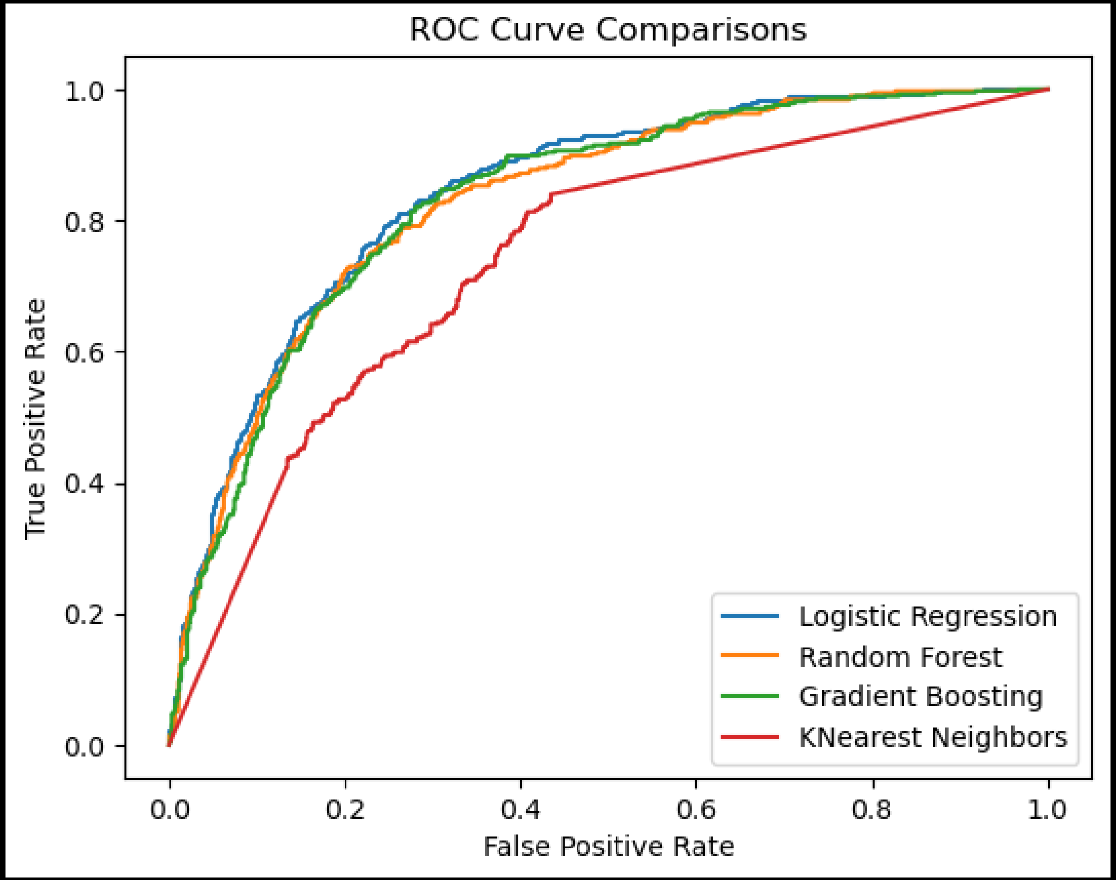


The classification report for the base logistic regression model:

	precision	recall	f1-score	support
Active Customers	0.91	0.74	0.82	1033
Churned Customers	0.53	0.81	0.64	374
accuracy			0.76	1407
macro avg	0.72	0.77	0.73	1407
weighted avg	0.81	0.76	0.77	1407

Classification Report for logistic regression model

	precision	recall	f1-score	support
Active Customer	0.91	0.74	0.82	1033
Churned Customer	0.53	0.80	0.64	374
accuracy			0.76	1407
macro avg	0.72	0.77	0.73	1407
weighted avg	0.81	0.76	0.77	1407



Modeling: Comparative ROC

- Due to computational resource constraints, probabilities for the SVC were not calculated
- Except for KNN, performance is relatively comparable for the all models
- Logistic regression was chosen based on both recall score and it is computationally least expensive

Modeling: potential enhancements

- Hyperparameter tuning didn't increase model effectiveness
- Modification to preprocessing steps
  - Filtering features based on influence
  - Use a different scaler, such as sklearn's MinMaxScaler vs StandardScaler or a combination based on each feature
  - Leveraging different resampling methods (Random over sampling or ADASYN)

Business recommendations: service bundles

- During EDA, noted several services that were negatively correlated to churn
  - Tech Support
  - Online Security
  - Online Back-ups
- Bundling these services with others may help with customer retention

Business recommendations: potential internal issues

- Further research is needed for customers receiving paperless billing, as they appeared to churn more than not
  - Electronic invoices may be getting caught in customer spam filters
  - Need to look at business email sender reputation scores
- Customer leveraging electronic check as a payment method were also more likely to churn vs those paying by paper check, automatic bank transfers or credit cards
  - Potential internal systemic issue that is causing frustration for consumers

Thank you!