Assignment 1: Predicting Customer Churn

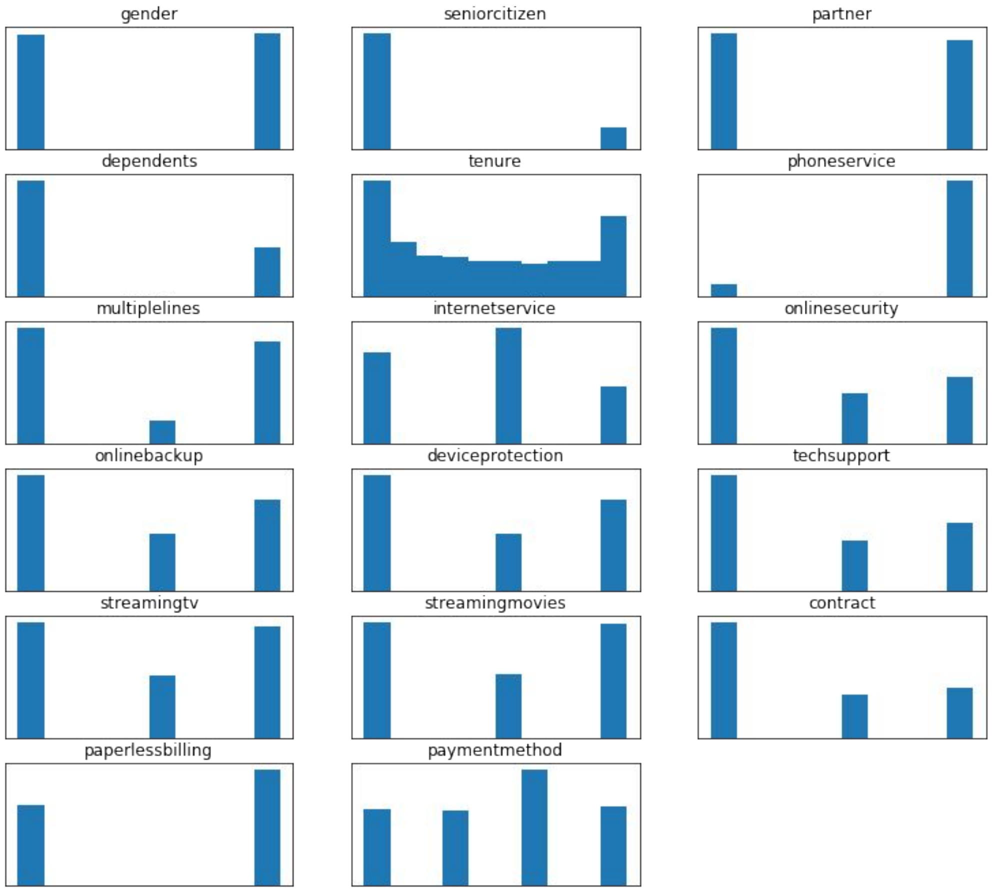
Irene Wei (zw575)
Zhiling Hu (zh347)

Data Exploration

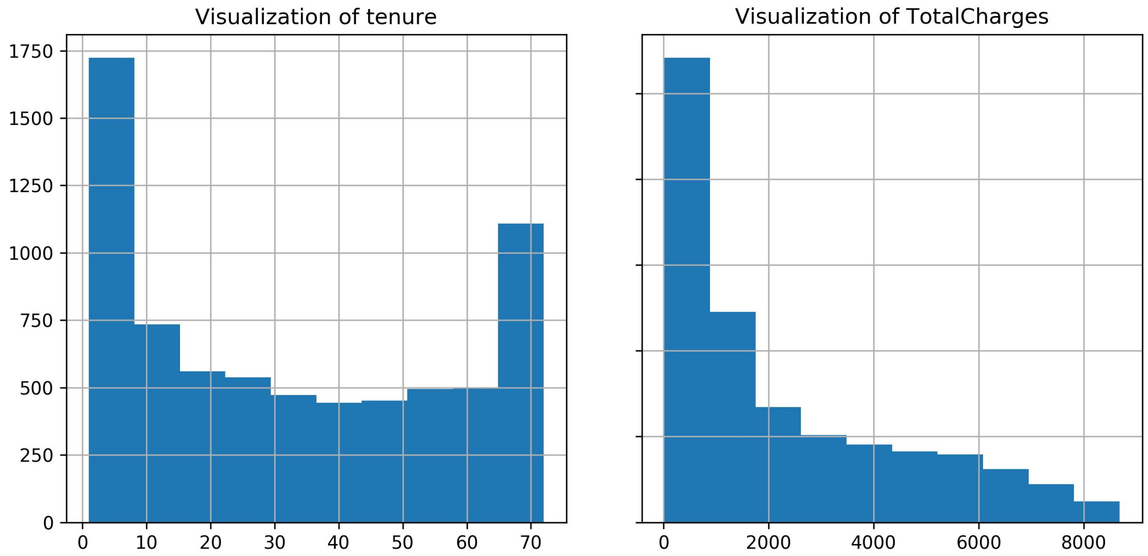
- The raw data contains 7043 rows (customers) and 21 columns (features). The “Churn” column is our target.
- There are 11 missing values in “TotalCharges” column, and we removed all rows with missing values since the amount is negligible.
- For columns that have two values, Yes and No, we performed one-hot encoding by replacing numerical values, 1s and 0s, with Yes and No.

Visualization of the data - Plot 1

- To get a sense the distribution of data for each attribute.



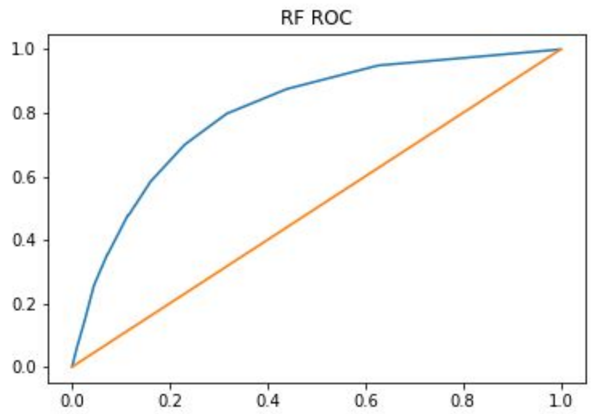
Visualization of the data - Plot 2



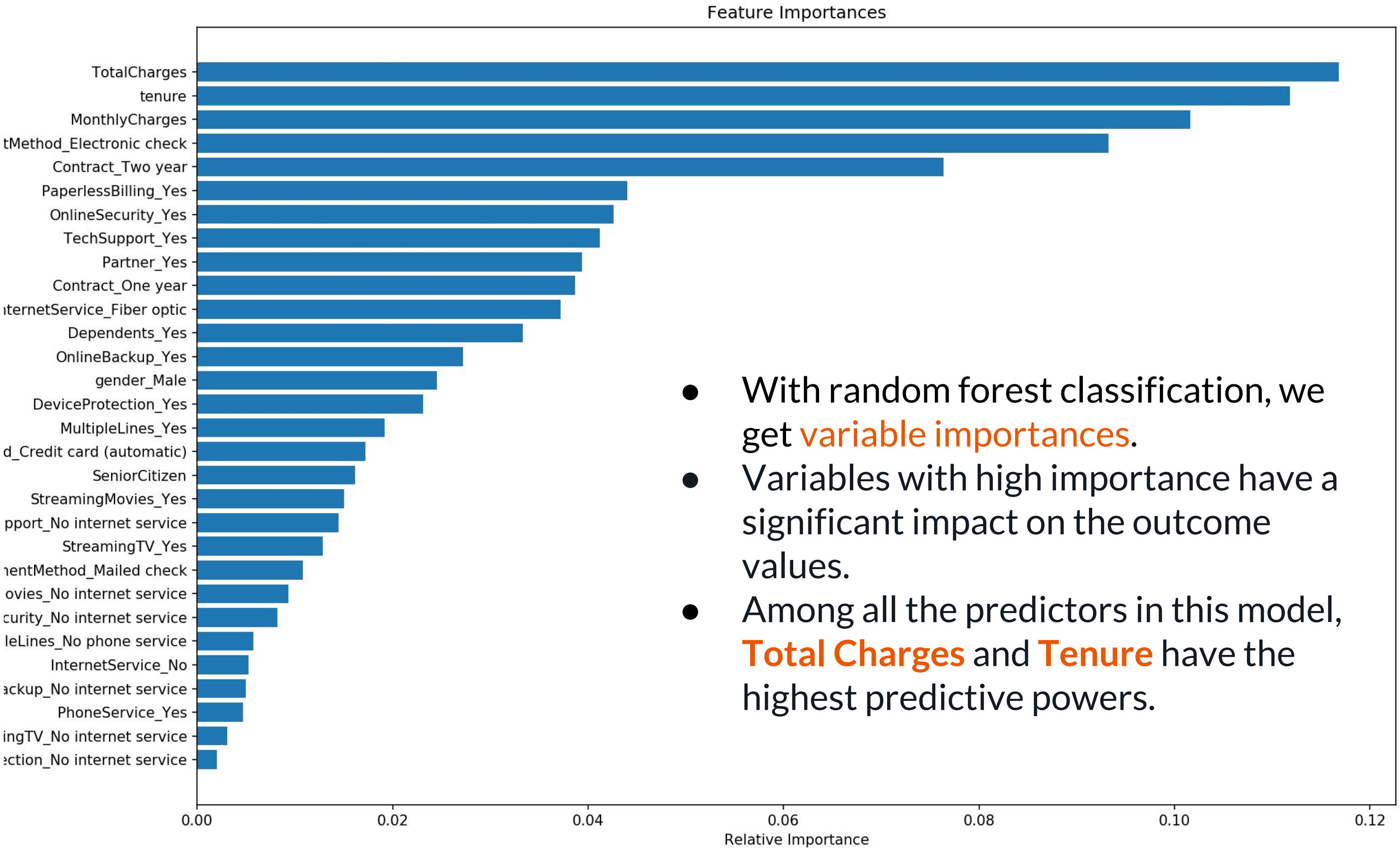
Histograms of TotalCharges and tenure.

Numeric features like tenure, total charges can generalize a group (or cohort). We tried to identify a group born in a certain timeframe.

Best Method



- Best Model: Random Forest Classification
- Details:
    - Split data to train set and test set
    - By calculating the percentage of “Yes” and “No” in churn column, identify whether data is imbalanced
    - Use SMOTE for oversampling
    - Among 3 different models we try, Random Forest and Logistic Regression are better in scores
- Select Random Forest for further analysis Accuracy with Random Forest: 0.845593
- Mean cross validation score: 0.828476



- With random forest classification, we get variable importances.
- Variables with high importance have a significant impact on the outcome values.
- Among all the predictors in this model, **Total Charges** and **Tenure** have the highest predictive powers.

Recommendations

- Customers with lower tenure are more likely to terminate the contract. We suggest to target customers with less than 12 month tenure.
- As total charges rise up, people feel more pressured to leave the program. We suggest to find a threshold/ trade-off between profit earned and the number of customers.
- The other factors that could be taken into account are contract plans and payment methods. We had a rough guess that customers with long-term contracts, such as one and two year's, are much less likely to churn than the monthly ones. We suggest to promote customer to switch to long term contracts.