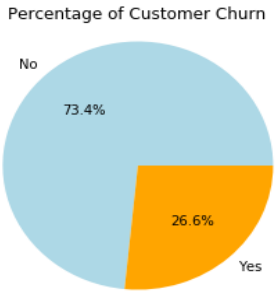
BT 2101 Homework 2: Predicting Customer Churn

1. Data Cleaning

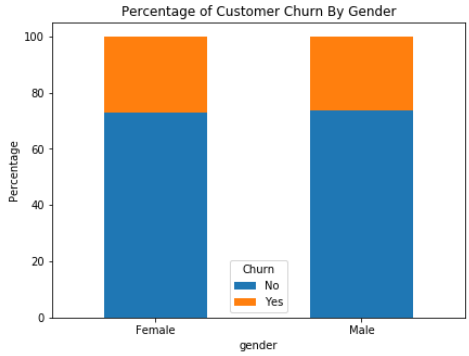
11 rows of data with empty cells are removed. For 6 attributes, “OnlineSecurity”, “OnlineBackup”, “DeviceProtection”, “TechSupport”, “StreamingTV”, “StreamingMovies”, cells filled with “No internet service”, was replaced with “No”. For “MultipleLines”, we treat “No phone service” as “No”.

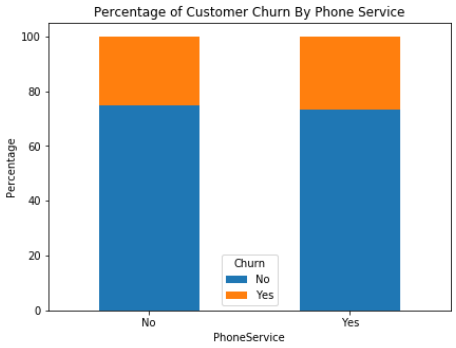
2. Exploratory Data Analysis

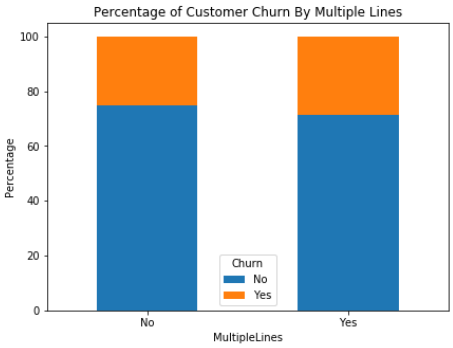


Only 26.6% of customers decided to churn, while 73.4% did not churn.

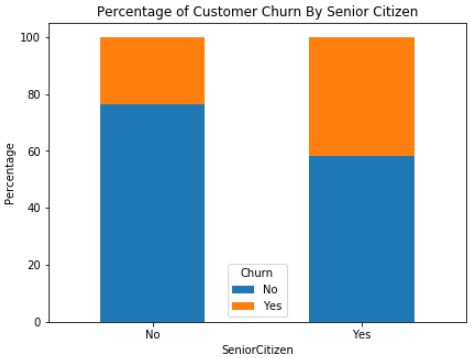
2.1 Categorical Variables

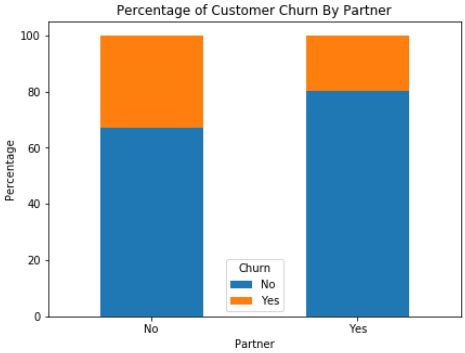


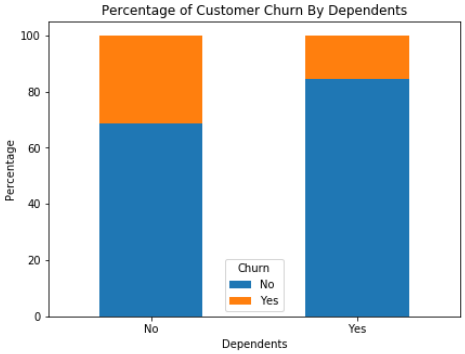




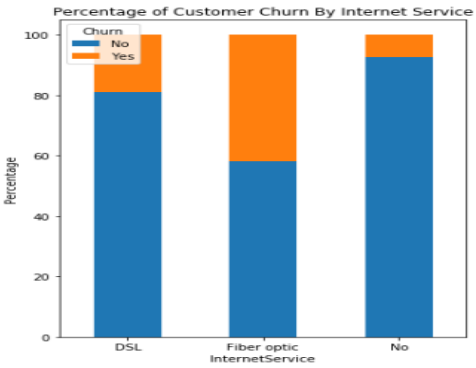
For attributes Gender, Phone Service and Multiple Lines, there are no significant differences in the churn rate.

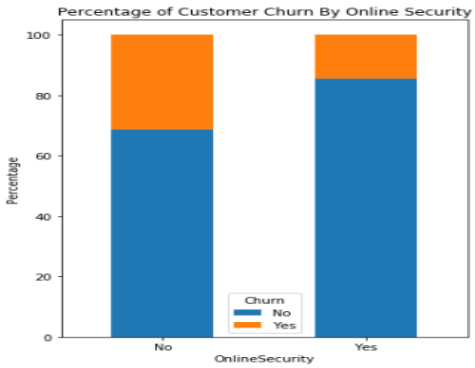


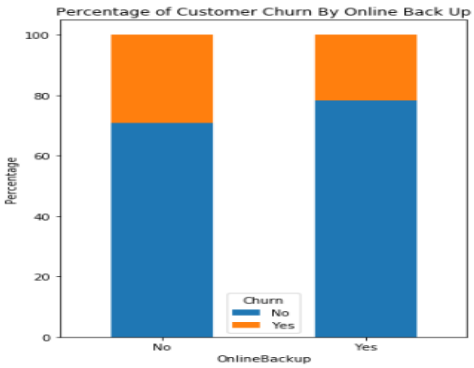




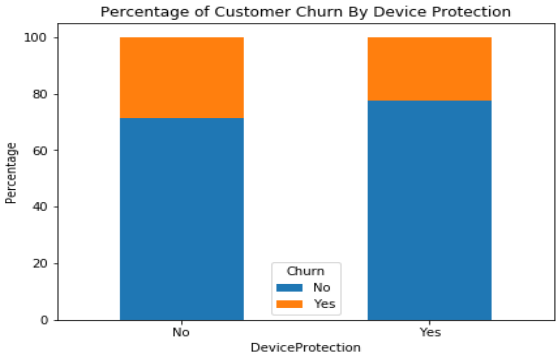
Churn rate is higher among senior citizens than those who are not. Customers with partners and dependents have lower churn rates compared to those without any partners or dependents.

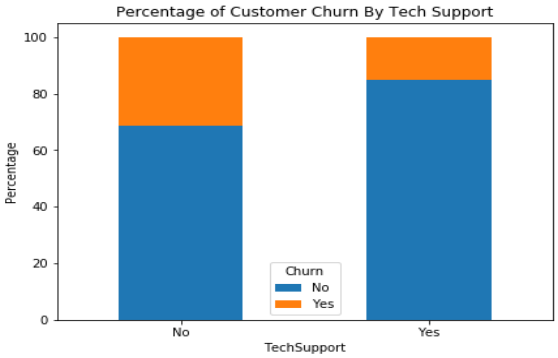




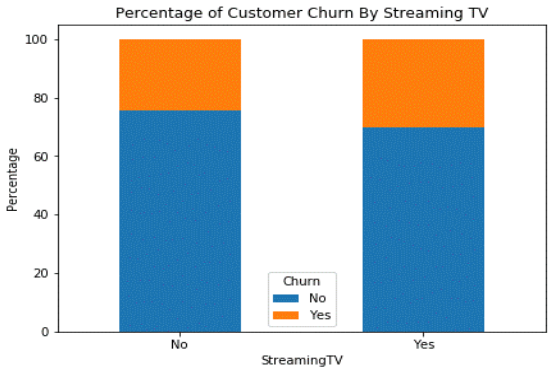


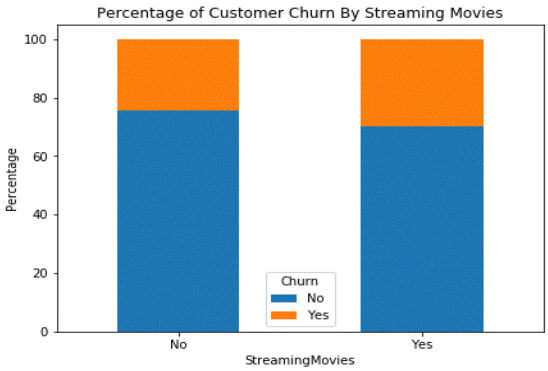
Churn rate is much higher for those with fiber optic internet service, followed by DSL and then those without any internet service. More customers without online security and online backup churned compared to customers with these services.



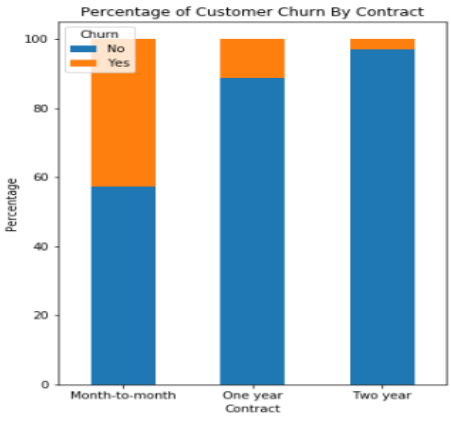


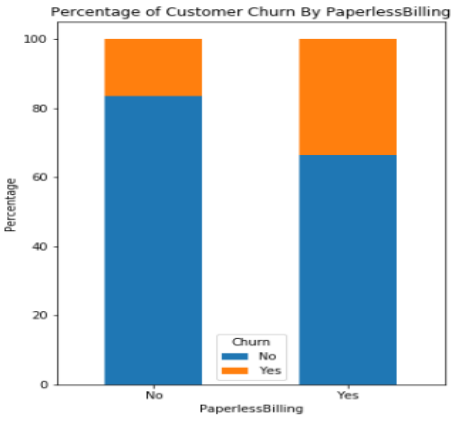
Customers with device protection and tech support have lower churn rates than those without.

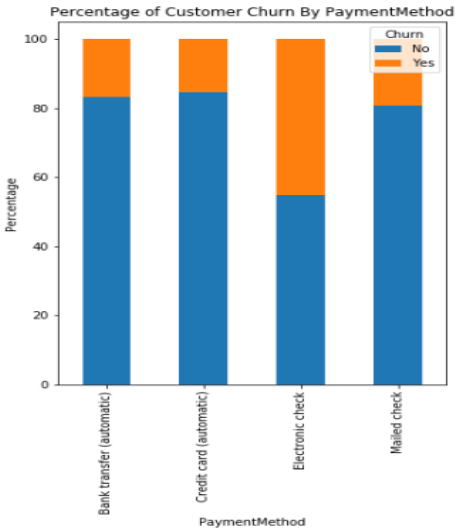




Customers with streaming TV and streaming movies have a higher churn rate than those without.



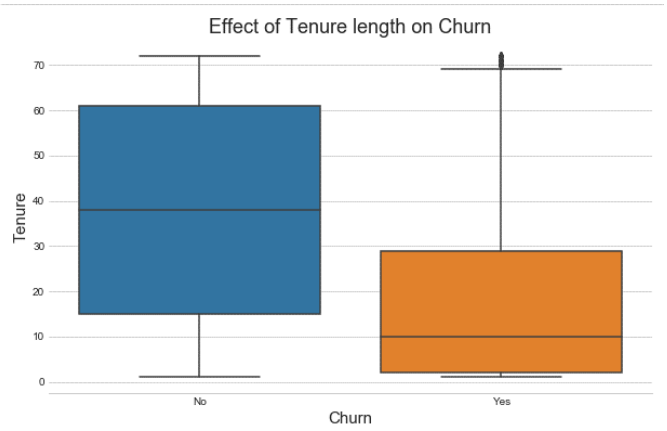


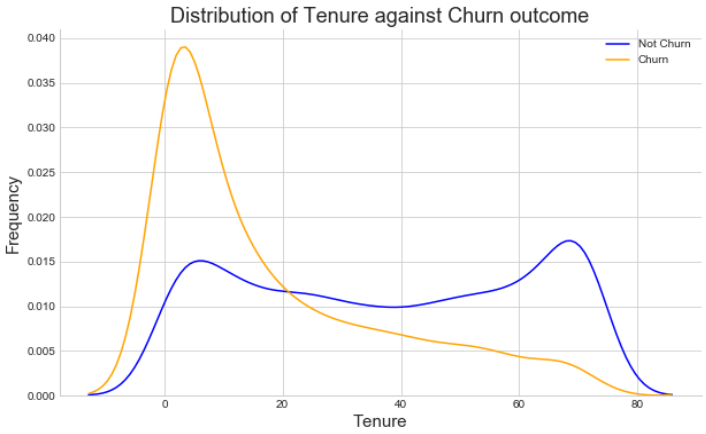


Few customers with 2 year contracts churn. Churn rate is higher when customer's contract is 1 year and is highest for month-to-month contracts. Churn rate is lower amongst those without

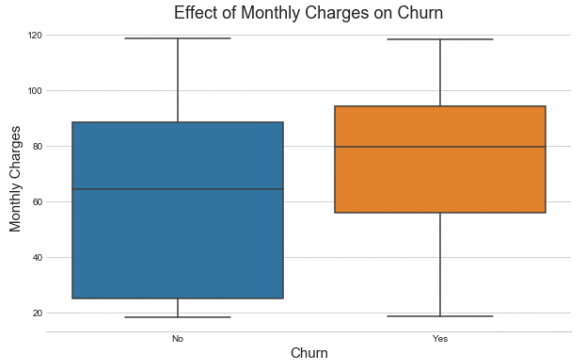
paperless billing compared to those with paperless billing. Bank transfer(automatic), credit card(automatic), and mailed check payment methods all have similar churn rates. Those paying via electronic check have a much higher churn rate.

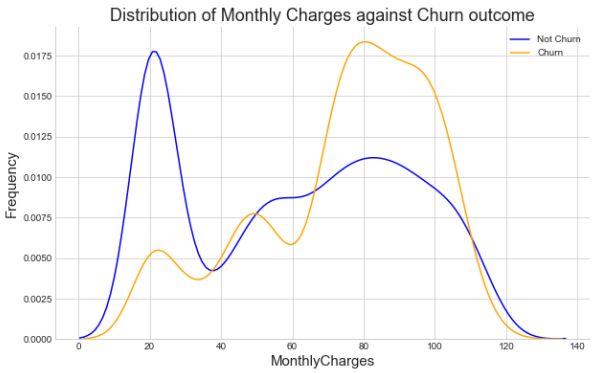
2.2 Continuous Variables



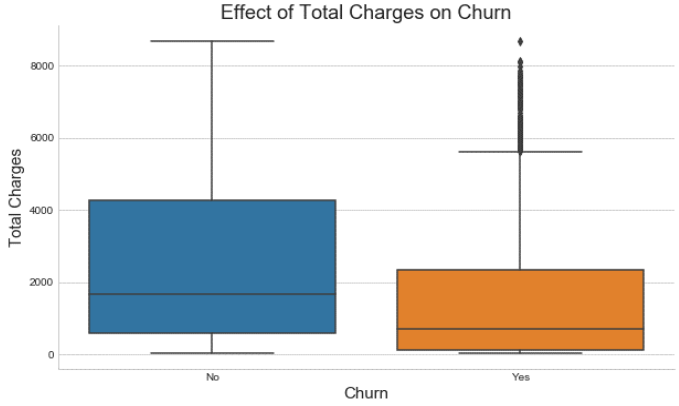


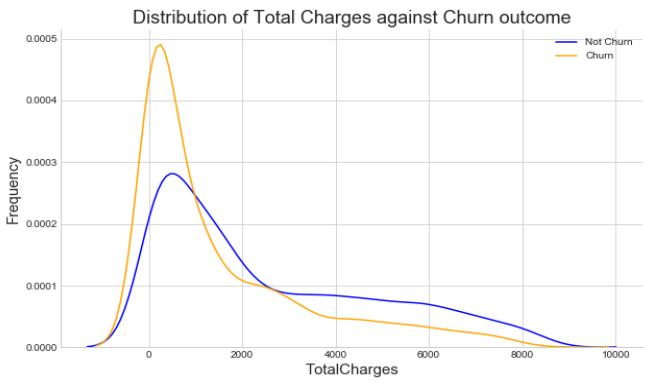
Median tenure of customers who churned is 10 months, while non-churn customers' was higher at around 40. Those who churned have tenure periods of 0 to 20. Those who did not churn usually have tenure period between 0-20 months or above 60 months.





Customers who churned have a higher median monthly charge at 80 while non-churn customers' median was around 60. Customers who churned have a relatively high monthly charge of 80 to 100. Non-churn customers have a lower monthly charge of 20 to 40.





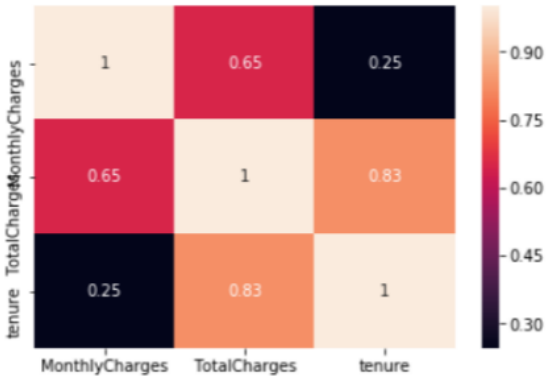
Customers who churned had a higher median total charge while those who did not have a lower median total charge. The value of total charges does not seem to affect churn rate as there was a similar proportion of customers who churn and do not churn regardless of the total charges.

3. Feature Importance

To decide on the importance of each feature, Chi-square test is carried out.

```
Features with p-values that are greater than 0.05:
PhoneService: 0.3514409861316786
gender: 0.6107282754601306
```

Phone service and gender have p-values greater than 0.05, thus they are insignificant and are removed.



Total charges is highly correlated (greater than 0.5) to tenure and monthly charges, thus it is redundant and is removed. Total of 3 features are removed.

4. Data Modelling

Accuracy score and cross validation score is used to evaluate models. For accuracy scores, GridSearchCV is used to get optimum value of parameters to reduce overfitting caused by using default values. Cross-validation score creates different subsets of the data to be used at least once for training and for testing to train and evaluate the model, reducing target leakage in data.

4.1 Baseline Model (Zero Rule Algorithm)

This model is used as a point of reference to compare to the other models. Accuracy is 72.65%

4.2 Logistic Regression

L1 Regularization carries out feature selection while L2 Regularization does not. Since feature selection was already done, L2 regularization will be used. Accuracy and cross validation score for this model was 80.569% and 80.204% respectively.

4.3 Decision Tree

Pre-pruning by finding optimal depth of the tree and criterion was done with GridSearchCV to prevent overfitting. Optimal features and accuracy score are shown:

```
{'criterion': 'entropy', 'max_depth': 3}
Accuracy of Decision Tree Classifier: 78.53080568720378%
```

Cross validation score is 73.35%. However, one decision tree might not be as accurate as advanced models like ensemble learning methods as it might have large variance and biasness. Thus, ensemble methods are explored below.

4.4 Bagging Model

For this model, accuracy and cross validation score was 79.14% and 78.07% respectively. Bagging reduces variance and avoids overfitting by creating separate decision trees with different training sets and averaging resulting prediction. However, most trees generated in this method use the strongest feature among all features to do splitting, increasing correlation among trees and thus biasness. To decorrelate the trees, random forest is used.

4.5 Random Forest Model

A random subset of features is selected from all features during each split, and the best feature within the subset is used to do the splitting. Due to combination of bootstrap samples and random draws of features rather than choosing the strongest feature, the trees are less correlated, reducing bias. Optimum parameters and accuracy score were:

```
{'criterion': 'entropy', 'max_depth': 8, 'n_estimators': 600}
Accuracy of Random Forest: 80.61611374407583%
```

Cross validation score was 77.97%.

4.6 Adaboost

Adaboost averages several weak learners and combine them into one strong learner to carry out predictions. Each weak learner is built based on performance of previous learner, thus eventually the strong learner has a lower bias and variance than each weak learner, improving prediction. Accuracy score and cross validation score was 80.90% and 80.22% respectively.

4.7 Model Comparisons

	Model	Accuracy Score	Cross Validation Score
0	Baseline Model	72.654028	-
1	Logistic Regression Model	80.568720	80.2046
2	Decision Tree Classifier (Entropy)	78.530806	73.3503
3	Bagging Model	79.146919	78.0716
4	Random Forest Model	80.616114	77.9721
5	Adaboost Model	80.900474	80.2189

Adaboost had the highest accuracy score and cross validation score. From ROC curve(Appendix Figure 1), Adaboost's area under curve is highest. Thus, Adaboost is the most appropriate model.

5. Limitations

The data only showed whether the customers were senior citizens or not. However, people in different age groups might have subscribed to different services (e.g. streaming TV), different type of plans and thus have different churn rates. Past years' data could also be given to see if there are any trends in the attributes that led to changing churn rates and allow for possible forecasting of future churn rates in current customers, which helps the telco to have a better idea of their future business model to retain and attract more customers.

Appendix

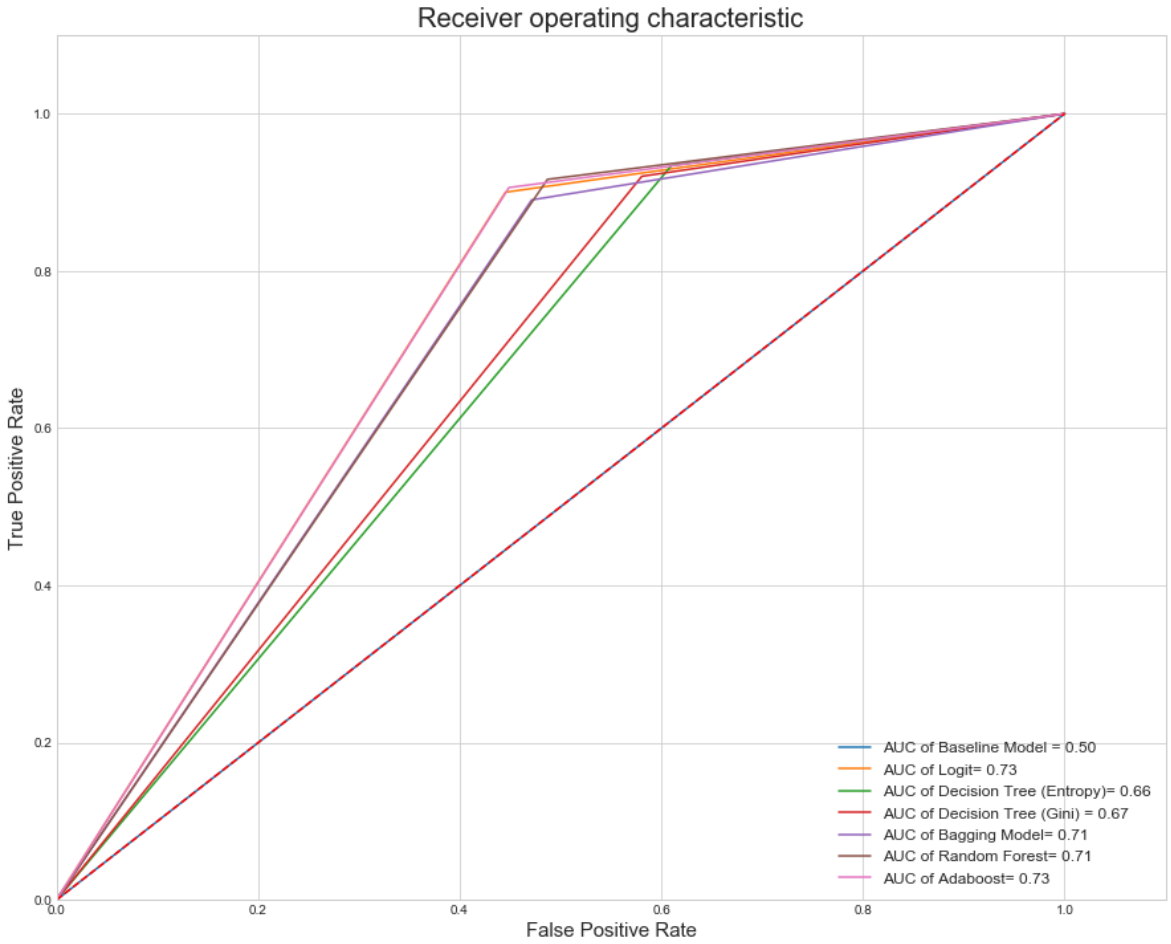


Figure 1