BT2101 Homework 2 **A0177660X**

Assignment Description:

The data set in this assignment is Customer Churn spreadsheet, which uses several attributes to classify and predict whether a customer will churn from the service or not. In this document, I will illustrate in detail how I apply machine learning algorithm and relevant concepts to the dataset as well as the overall model evaluation.

Step 1: Data Preparation

Data Cleaning:

- Drop the 'customerid' column as it's irrelevant to our prediction
- Reshuffle the ordering of columns so that `Churn` appears in the front
- Modify binary variables to make them with value 0 and 1
- Create dummy variables for categorical columns with more than 2 levels

EDA: [see appendix for according visualizations]

- For continuous variable:
    I. Plot boxplot and histogram to check extremes and distribution
    II. Plot correlation matrix and Kernel Density Estimation
- For categorical variable:
    III. Use bar plot to find the proportion of churning for each value inside.

Data Discretization:

- Categorize continuous variables and make them ready for classification models

Step 2: Features selection

Initial selection:

Drop columns without prediction power:

- Phone Service has no prediction power

By careful examination, we can find that whenever the phone service has a value 'No', the corresponding value in Multiple lines is always 'No Phone Service'. This means that phone service has no prediction power and we can drop it from the table.

Drop correlated columns:

- 'Tenure' * 'Monthly charges' is strongly correlated with Totally charges

Since 'Tenure'* 'Monthly Charges' approximately equals to Total charges (See Appendix), we can further simplify the model and avoid redundancy by dropping both Tenure and Monthly Charges columns. Then use 'TotalCharges' as a representative.

Secondary Data-driven Selection:

A train-test-split logistic regression model, with L1 normalization, is conducted for all the features available after initial selection. I checked the coefficients of each variable and set a threshold value of 0.07, which means variables with a coefficient between -0.07 and 0.07 are considered to have very low prediction power and dropped from the data frame. These variables are: \`Gender\`, \`Online Backup\`, \`Online Security\` and \`Partner\`.

Step 3: Model building and comparison:

[Check appendix for model accuracy]

Remarks on building models:

- L1 regularization is used, which adds a penalty term on the coefficients to restrict them from growing too large and as a result, lower the possibility that the logistic regression model over-fits the training sample
- 10-fold cross validation is used, which allows building models with more data and thus, helps identify and address overfitting
- Ensemble learning, which contains a set of homogeneous classifiers, are used and expected to produce better accuracy than a single classifier.
- Initially, all the attributes are used for each model. Then apply the updated dataset with selected columns to all the models, except random forest as it imposes an attribute-selection process for each individual classifier inside
- Confusion matrix, ROC and AUC curves are used to assess the model performance
- Use the method of oversampling on the dataset to address the problem of rare outputs. This method is used on CART, random forest as well as bagging (as they impose a majority voting mechanism). It turns out that oversampling improves the model performance dramatically (from 0.73 to 0.88).

Step 4: Model review and reflection:

Overall performance:

The average predictive accuracy of the model is around 0.85, a good indicator to show that the model has a satisfactory performance with a low chance of overfitting.

Moreover, after narrowing down the features, the classifier gets simpler without sacrificing accuracy (Some even have a small scale of improvement in terms of accuracy). By the principle of Occam's razor, the overall model performance is improved after dropping those columns. Furthermore, among all the models used, random forest with oversampled data gives the highest predictive accuracy of 0.90.

Some more data needed, if possible:

1. More demographics of customers, including income, education, location etc.
2. More data on customer behaviors, e.g. Time spent on the services, number of times of complaints……

Limitations (things could be improved):

1. For attributes `OnlineSecurity`, `OnlineBackup`, `DeviceProtection`, `TechSupport`, `StreamingTV` and `StreamingMovies`, I directly change its value to ‘No’ when its actual value is ‘No internet service’, without further investigation. It improves model performance but at the same time, introduces noise to the data.
2. If given more time, I would do additional train-test-split procedures on each model to further check whether the model over-fits the training data. If testing accuracy does not differ much from training accuracy, we are then confident to believe that the model does not over-fits the training data. Faced with limited time, I only apply this process once to a simple Logistic Regression model.
3. The model could be possibly improved by including a parameter optimization process. These parameters include step size, maximum iteration etc.

<u>References</u>:

[1] https://www.ibm.com/

[2] https://www.kaggle.com/learn/machine-learning

[3] https://elitedatascience.com/overfitting-in-machine-learning

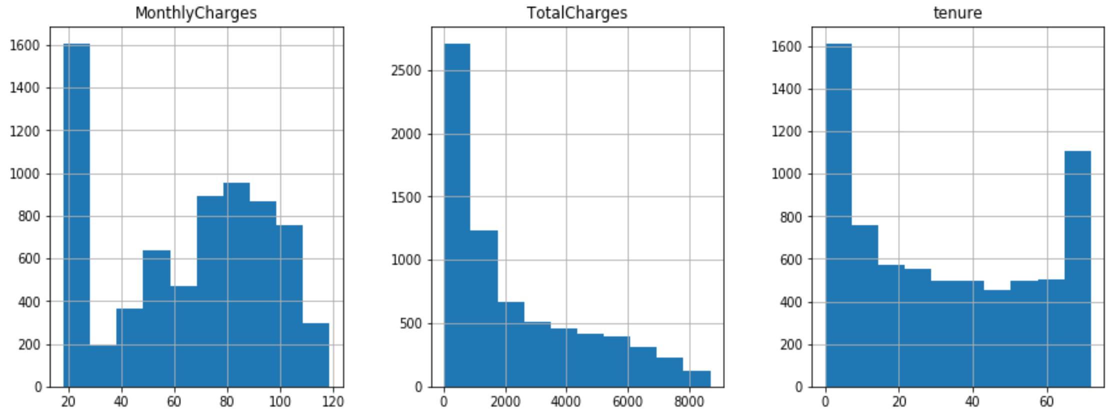
[4] Max Bramer, Ensemble Classifier. In: Ian Mackie(eds.) Principles of Data Mining, Third Edition; 2016. p. 209-230

[5] Peter Harrington, 2012. Machine Learning in Action. Shelter Island, NY: Manning Publications Co.

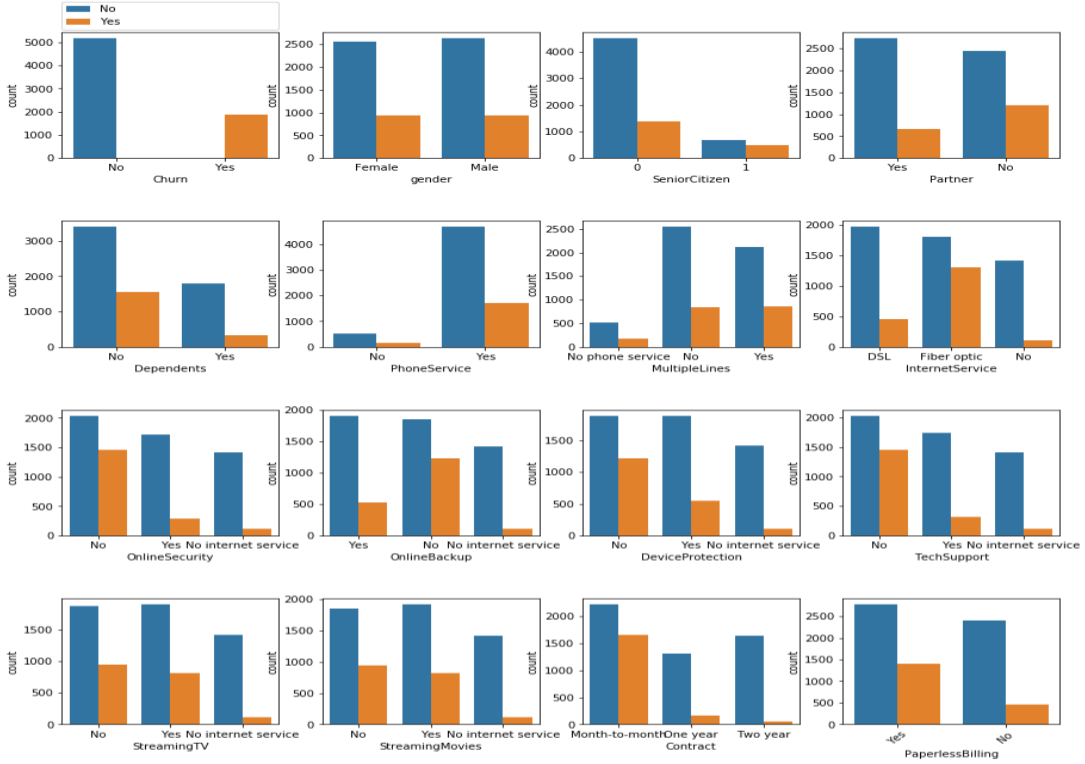
[6] https://github.com/mozartkun/BT2101_Tutorials_2018_2019_SEM1

Appendices:

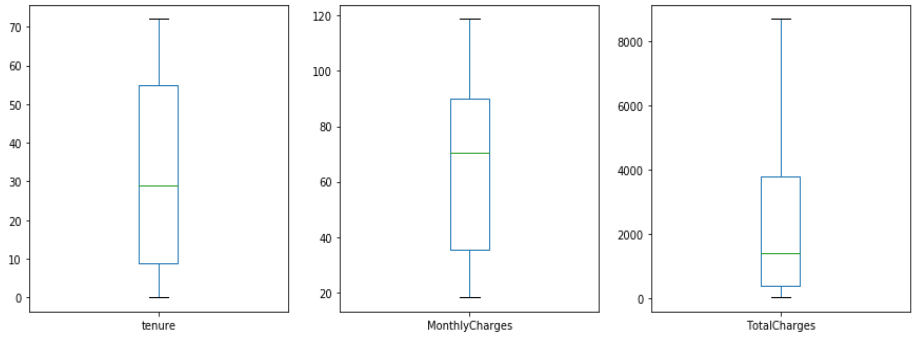
Distribution of Continuous Variable



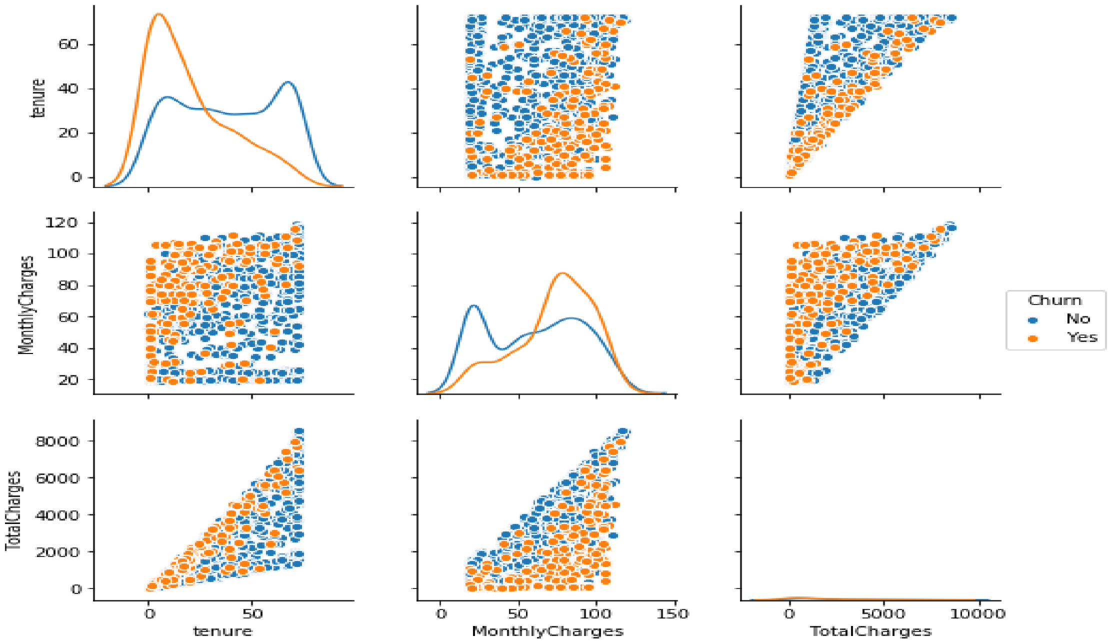
The proportion of 'Churn' for categorical variables



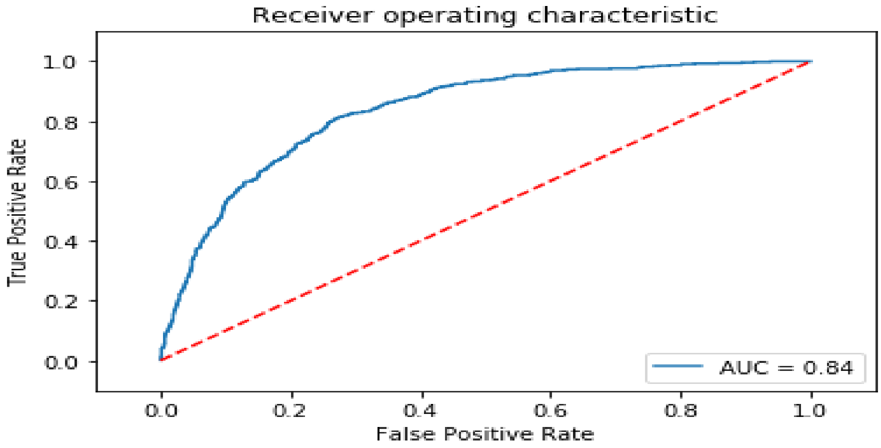
Box plot of continuous variable



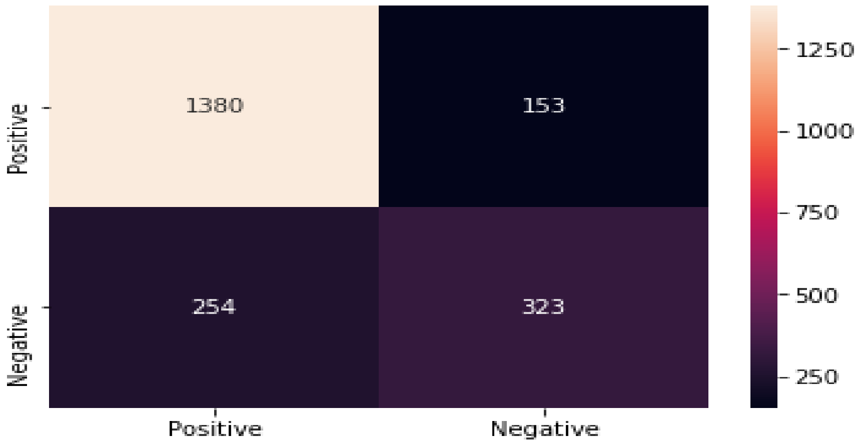
Kernel Density Estimation



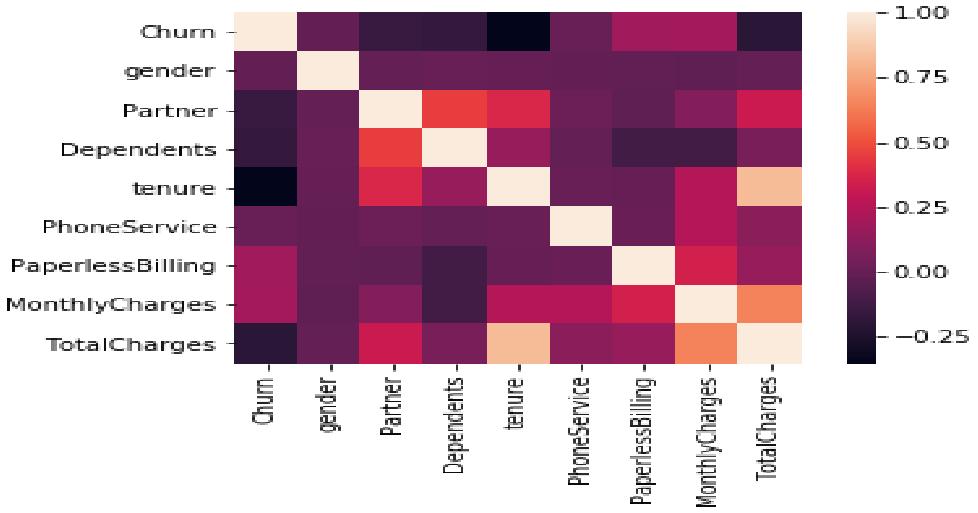
ROC and AUC curve for logistic regression



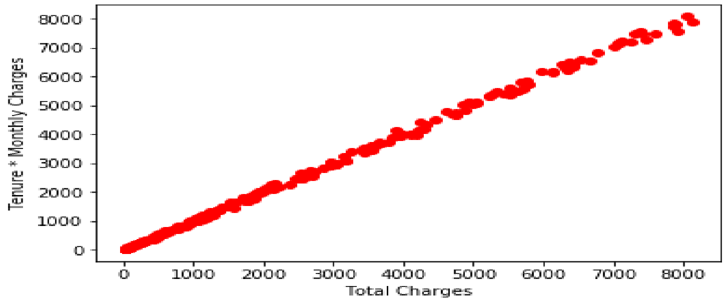
Confusion Matrix:



Correlation Matrix



Scatter plot of Total Charges and (Tenure * Monthly Charges)



Final Predictive Accuracy

Model	Model Accuracy
Logistic regression	0.805
Logistic Regression (selected features)	0.805
CART (all features)	0.738
CART (all features, oversampled)	0.885
CART (selected features, oversampled)	0.885
CART (Gini index)	0.885
Random Forest	0.791
Random Forest (oversampled)	0.909
Bagging	0.905
Bagging (selected features)	0.895
Ada-boosting	0.804
Ada-boosting (selected features)	0.806