Customer Churn in Subscription Business Model—Predictive Analytics on Customer Churn

Boyuan Zhang*

Department of Economics and Mathematics, New York University, New York, United States

*Corresponding author: bz2059@nyu.edu

Abstract. There is a growing tendency for more companies to develop towards a subscription business model. Under such a trend, it is important to learn about the customer churn rate within the business, learn from it and adjust business strategies accordingly. This paper aims to predict customer churn rate in subscription business models using a variety of machine learning algorithms. Through comparing the results from the different algorithms, the best algorithms can be identified so that it provides an insight on which algorithm a subscription business should choose in order to predict customer churn most effectively. In this work, a total of 21 features and 9 algorithms are taken into account. Through a set of rigorous procedure including data preparation, feature engineering, feature selection, model building, and finally, model evaluation, three algorithms, namely Logistic Regression, Gradient Boosting (SMOTE) and Neural Network outperformed other 6 algorithms. The best performing algorithm being Logistic Regression with its 79.6% prediction accuracy, thus the conclusion that when subscription business predicts customer churn rate, Logistic Regression is the most preferable algorithm. During the process of feature engineering, SMOTE did not improve the model performance as it supposed to, so it is not recommended during the model building process.

Keywords: customer churn prediction; machine learning; deep learning.

1. Introduction

Customer attrition, often known as customer churn, is the tendency of customers to discontinue being clients of a specific company, particularly a subscription firm. It actively indicates a firm's performance on how well they are doing in terms of selling their products or services. Nowadays, the number of businesses that use a subscription business model, which is based on customers paying a recurring amount on a regular basis in exchange for a service or the use of a product, has been increasing. Not only does traditional businesses like newspaper and magazines use a subscription business model, but rising streaming sector like Netflix and Spotify as well. Even cars and airlines move to subscription models. Under such circumstances, high retention rate of the customers is crucial for a subscription business's survival. As a result, customer churn is a scourge that all subscription businesses want to avoid and minimize its impact. The importance of predicting how and when the customers are likely to churn is evident. Businesses need to understand how churn impacts their revenue goals and make predictions about how churn rate is going to fluctuate in the future before taking actions accordingly.

There are a few existing research endeavors on the topic. Reviews of early research on customer churn reveal the fact that the bulk of the research either focus only on one or few algorithm models to predict potential customer churn, or the initial features selected were not comprehensive enough to predict the overall churn rate. Owczarczuk came to the algorithms of logistic regression and decision tree for predicting the churn rate [1]; Irpan used a neural network algorithm model for deployment modeling to explore potential lost customers [2]; The algorithm model Li et al proposed to predict customer loss is based on Improved Bat Algorithm (IBA) and optimized Extreme Learning Machine (ELM) [3]. Once potential customer loss is found, telecom companies will actively communicate with customers and find ways to retain them. The research that applied the most algorithms is by Umayaparvathi et al, with a total of 7 of them including Gradient Boosting [4], Decision Tree [5], Support Vector Machine [6], Random Forest [7], K-Means [8], Ridge Regression [9], and Logistic Regression [1]. However, the features examined are only customers' behavioral information, customer care and demographics. On the other hand, the research with a comparatively more comprehensive features collected by Ismail and Mohammad et al, which includes demographic, billing data, usage pattern and customer relationship, only focused on Neural Network and Regression [1]. Under such circumstances, this research aims to use a variety of algorithms based on most of the features related to predict customer churn, which fills the gap that previous research has on this topic.

The purpose of this research concerning a total of 21 features ranging from demographic, billing data, usage pattern, to relationships is to provide a thorough analysis using a variety of algorithms, before finally comparing the prediction results to find which one of the algorithms provides the best prediction results. Through a series of steps including data analysis, feature engineering and feature selection, a total number of 9 algorithms were used to build predictive churn models. Using the result predicted by the most accurate algorithm found by this research, subscription businesses like telephone service companies can have the idea on which algorithms is likely to produce the result that helps with the companies the most, then make proactive changes to their retention efforts that drive down churn rates and stem the flow of churned customers. Instead of seeking aimlessly between the algorithms, subscription businesses can stick to the one best performed algorithm, saving them a huge amount of time.

2. Methodology

2.1 Dataset preparation

In order to build predicative churn models, this paper explores the variables based on the dataset to have a general understanding about the data. The dataset is originally attained from Kaggle [10], with the type of task of classification (churn or not churn). The number of the data item is 7043 and the number of features is 21. The total customer attrition in data (shown on Fig. 1) shows the total number of churn customers is 1869, while the number of none churn customers is 5174. The corresponding overall churn rate calculated is 26.54%.

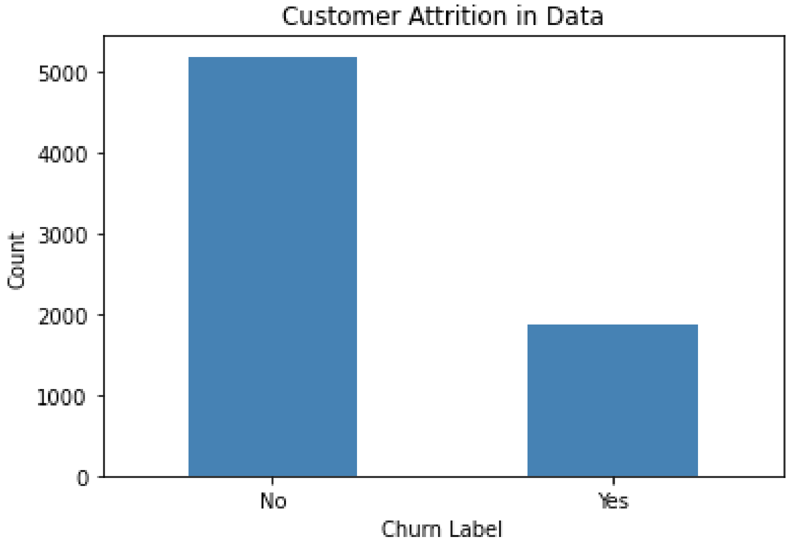


Figure 1. Total Customer Attrition

Taking a deeper look into the data, it is easily identifiable that the variables are divided into two categories: Categorical Variables and Numerical Variables. While there are a total of 16 categorical variables divided based on gender, having phone service or not, contract length and so on, there are only 3 numerical variables, being tenure, Monthlycharges, and Totalcharges. Based on the original variables, the variable distributions by attrition type are created to build a clear connection between the variables and customer attrition.

2.2 Feature engineering

After examining the current variables, the next step is feature engineering. Feature engineering is the second most important step of the predictive analytics pipeline, with its importance right after the framing of the problem. It refers to manipulation, which includes addition, deletion, combination, mutation of the data set to improve machine learning model training, creating special variables that are candidate inputs to models. The resulting special variables are more related to the target, thus a model with greater accuracy and better performance is more likely to be achieved.

Mathematical and statistical methods are used to create variables. Two more variables were first added. The first variable is created based on all the data with no Online Backup, no Device Protection and no Tech Support and named it “Non-Protection”. The second variable is the sum of all services and named “TotalServices”. a list is then used to calculate ratio for all the variables, named ratio variables. Resulting number of variables created is 33. Log of the numeric fields is also added, with the resulting number of 38 variables; Statistic features were then added, including the “mean”, the average of all numbers, “count”, the total number of data we have, “max”, the maximum among the data, “min”, the minimum among the data, “nunique”, unique values for each row of data, “std”, the standard deviation, “var”, the variance, “skew”, the skewness and “median”, the median. The resulting number of variables is 758; lastly a new variable named polynomial features is defined, a function is built and the resulting number of variables is 773.

Next, considering the categorical variables. The categories need to be encoded into numbers to go into a modeling algorithm. Target Encoding (for each possible category assign a value) and One-hot Encoding (produces one binary feature per category) are used to transform the categorical fields into numerical variables. The advantage of Target Encoding is that it does not add to the dimensionality of the dataset, which is among the biggest enemies of the modeling process. On the contrary, One-hot encoding is prone to create very high dimensionality depending on the number of categorical features. In the case of this particular dataset, both methods are applied for feature selection to figure out what’s best. The resulting total variables created by One-hot encoding is 32 and the total variables created by Target encoding is 21.

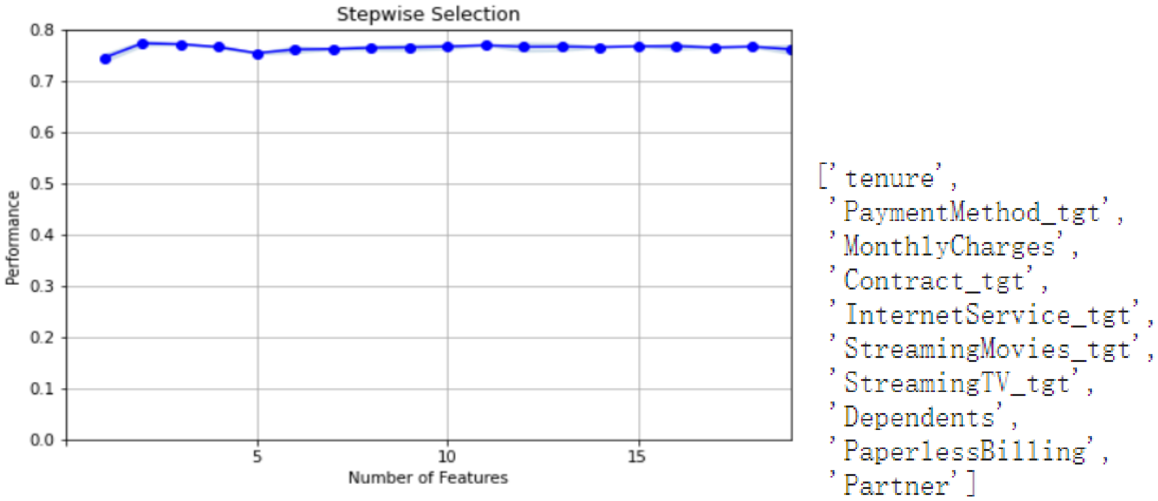
However, there is an imbalance between the goods and bads. SMOTE, or Synthetic Minority Oversampling Technique is used to upsample the bads so that the number of goods and bads matches. First randomly pick a point from the minority class, then determine the “k” and compute the k-nearest neighbors for this point. Finally, k new points somewhere between the chosen point and each of its neighbors is added. The model building algorithms can then to applied to build the model. In the notebook, a total of 9 algorithms are used including Logistic Regression, Decision Tree, K Nearest Neighbors, Random Forest, Gaussian Naïve Bayes, Light GBM, XGBoost, Gradient Boosting, and Neural Network.

2.3 Feature selection

After creating as many variables as possible and transfer them all into numerical values, the next step is to do Feature selection. According to Curse of Dimensionality, as dimensionality increases, data becomes sparse very quickly and all points become outliers, thus the difficulty to fit a nonlinear model increase. Consequently, the goal is to reduce dimensionality as much as possible and minimize the number of variables. Feature selection is the exact process where the dimensionality is minimized, or in other words, reducing the number of input variables which are non-informative or redundant. It would eventually result in a sorted list of the variables we need in the model. The resulting model based on the selected variables is going to be clear, accurate, and fast.

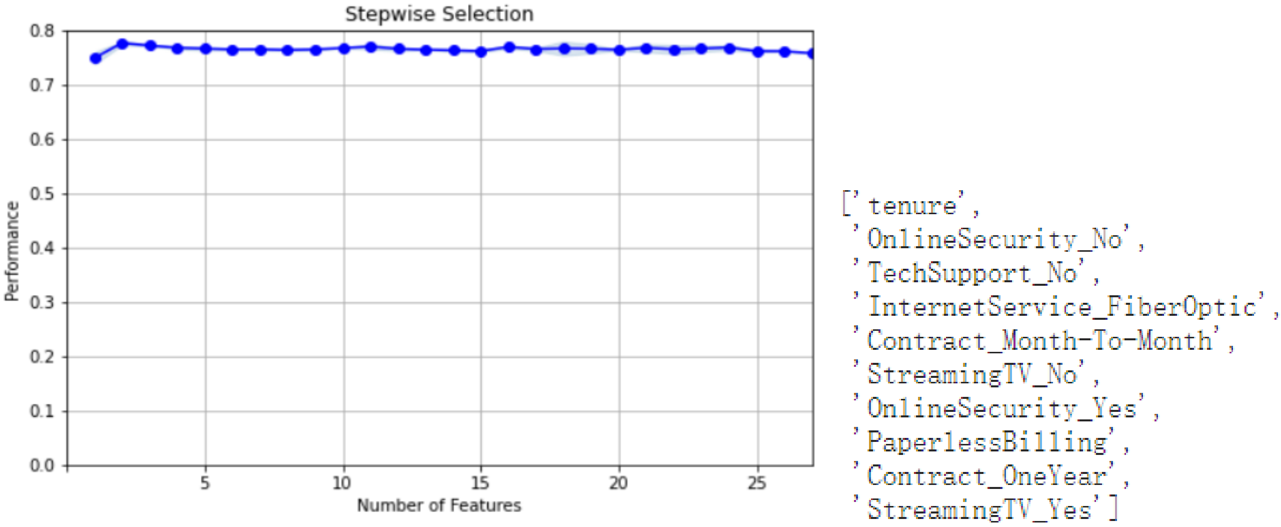
KS & Churn Detection Rate Univariate Filter and Cross-Validated Recursive Feature Elimination (Stepwise selection wrapper) are used for feature selection. Since filter looks at one variable at a time, the filter is firstly applied to the data since it is a faster way to quickly filter down the variable number. The goal of the filter is to know how important each variable by itself to predict y. Firstly, a function that calculate KS score and Churn Detection Rate for all features for both Target encoded data and One-hot encoded data is defined. The result is stored in a dataframe. A train test split is then performed on the encoded data. The features with average score of KS and CDR above 0.2 are selected. The result is 19 features selected for the Target encoded data and 27 features selected for the One-hot encoded data.

The variable number can be further reduced after using a stepwise selection wrapper. The goal is to create a short but conservative list of variables in the order of multivariate importance. Correlations are taken into account since one strong variable is chosen, another variable that is highly correlated with that variable is not going to add much to the performance. Wrapper allows to test on various number of input variables to see how the performance changes. As Fig. 2 and Fig. 3 shows, the performance peaked when the number of features is 2 in both the Target encoded data and the One-hot encoded data. However, the exact position of the best result is not always reliable, and as much as twice as many variables as what it says to be optimal can be kept. In this case, it is even more extreme. As many as 10 variables are kept in both datasets.



```
['tenure',
 'PaymentMethod_tgt',
 'MonthlyCharges',
 'Contract_tgt',
 'InternetService_tgt',
 'StreamingMovies_tgt',
 'StreamingTV_tgt',
 'Dependents',
 'PaperlessBilling',
 'Partner']
```

Figure 2. Target Encoded Variables Selected by Wrapper



```
['tenure',
 'OnlineSecurity_No',
 'TechSupport_No',
 'InternetService_FiberOptic',
 'Contract_Month-To-Month',
 'StreamingTV_No',
 'OnlineSecurity_Yes',
 'PaperlessBilling',
 'Contract_OneYear',
 'StreamingTV_Yes']
```

Figure 3. One-hot Encoded Variables Selected by Wrapper

2.4 Machine Learning Model Building

A model can subsequently be built after we have the final variable number gotten down to 10.

For each algorithm, the average k-fold cross-validated accuracy, roc-auc, and KS scores are calculated. If smote is True, each model will be additionally trained on upsampled train sets during cross validation.

3. Model Evaluation

Looking at the model performance illustration from Fig. 4 and Fig. 5, it can be easily seen that the overall pattern for one-hot encoded dataset and target encoded dataset looks very much alike. This indicates that both feature engineering method perform very similar on the baseline models with default hyperparameters. Logistic Regression is the best model for both datasets. One thing that requires special attention is that SMOTE does not always improve model performance. In fact, SMOTE rarely improve model performances since datapoints that do not originally exist were created and added to the model.

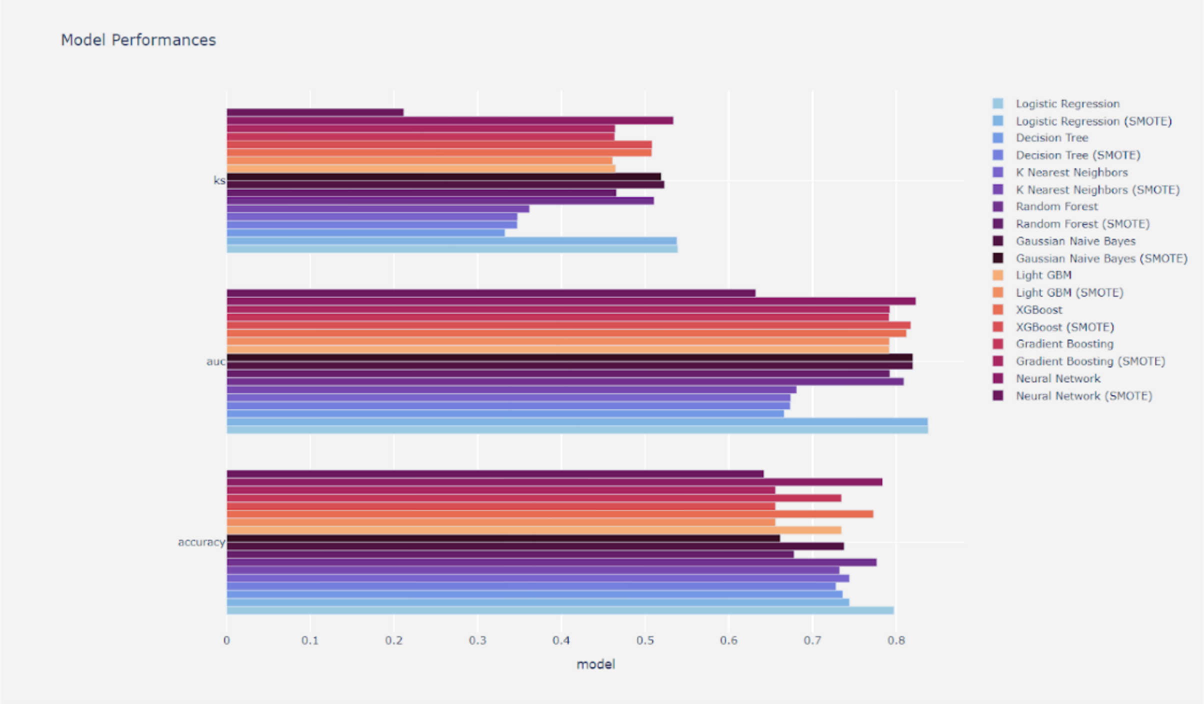


Figure 4. Model Performances for Target Encoded Data

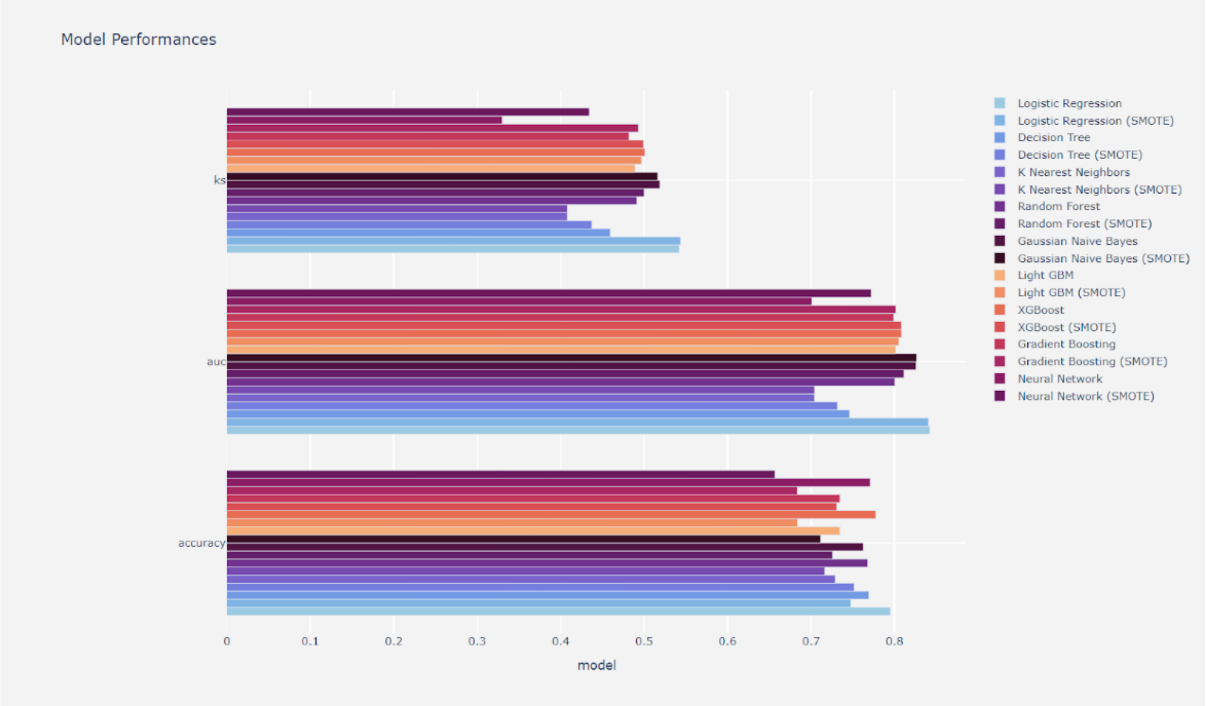


Figure 5. Model Performances for One-hot Encoded Data

According to the figures, the top 3 models are Logistic Regression, Gradient Boosting Tree and Neural Network. Model Tuning is then performed. Model Tuning is the process of maximizing model performance without overfitting or creating too high of a variance. It is accomplished by selecting the appropriate “hyperparameters”, which are the parameters that define the model architecture. The dataframe that shows the accuracy, the Roc-auc score and ks score of the three best performing algorithm and their corresponding visualization in a bar chart is shown in the following Table 1 and Fig. 6.

Table 1. Scores of the Three Best Performing Algorithms

Algorithm	accuracy	auc	ks
Logistic Regression	0.796238	0.837811	0.541777
Gradient Boosting (SMOTE)	0.773337	0.842570	0.550747
Neural Network	0.794282	0.837760	0.542858

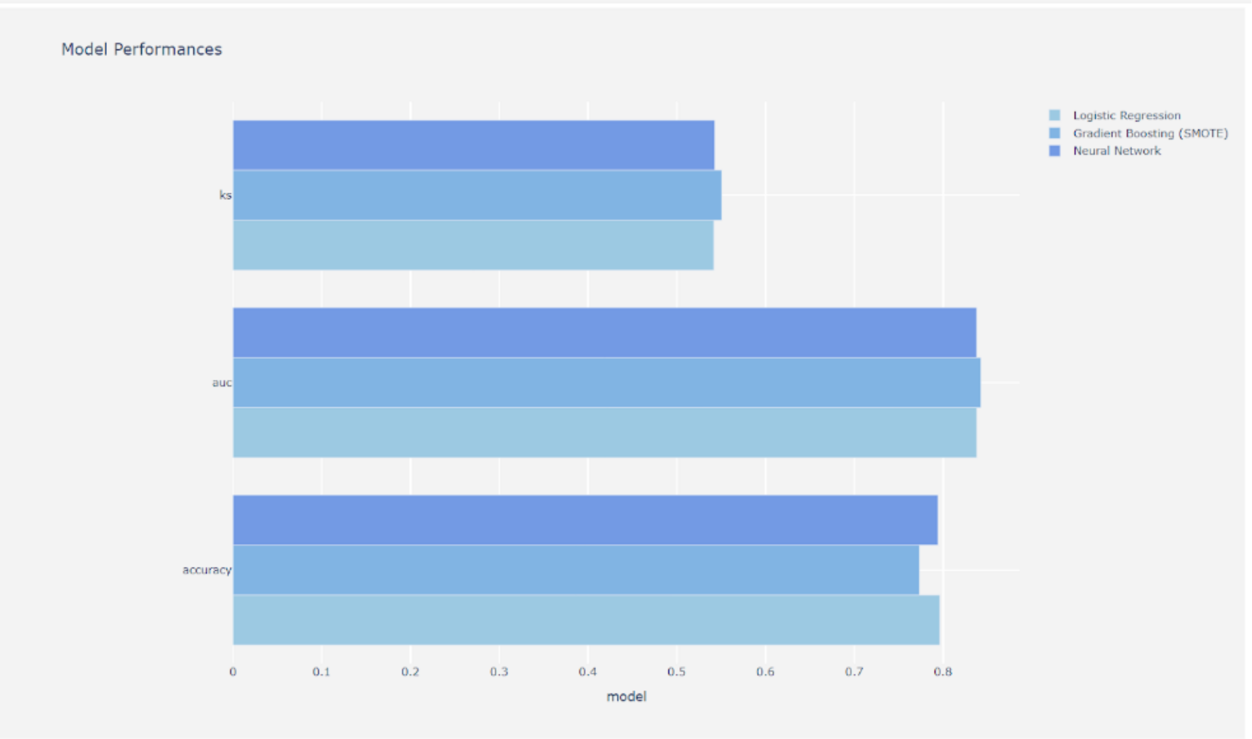


Figure 6. Bar Chart of the Three Best Performing Algorithms

4. Conclusion

Logistic Regression model, with an 79.6% prediction accuracy, outperformed all other models based on performance indicators. It is also among one of the simplest operating models. This gives us the insight that for later similar customer churn prediction problems, Logistic Regression is more likely to give us the best result. One thing about the notebook that has room for improvement is the SMOTE used during the model building stage. The result informed us that creating data points that are not originally in the dataset is not going to help that much. Instead, randomly downsample or undersample the goods could be used to balance the data sets. Since all of the data used are exactly the data points in the dataset but with random selection, a better performance than SMOTE could be expected.

Successful subscription businesses are based largely on recurring revenue made by subscriptions. Thus, customer experiences and relationships are the most important factors affecting these businesses. Based on the result of the models, a list of the predicted churn customers can be created.

Through analyzing their respective characteristics, a specialized marketing strategy could be made to prevent them from churning. A more frequent survey could be constructed during the customers' use of the service to understand what exactly are the customers not satisfied with, then companies could improve their services based on the result of the survey. With the predictive analysis, subscription companies are more likely to preserve former customers in designing specialized market retention programs.

Moreover, analyzing customer churn patterns provides insights on subscription companies' innovative opportunities. Based on the customer churn rate of the previous products or services, companies are able to tell which products or services are preferred by their customers as a whole, thus get a good grasp of the taste and preference of the general customers. When developing new products or services, they can learn from their past successes and adjust accordingly.

References

[1] Umayaparvathi, V., and K. Iyakutti. "A survey on customer churn prediction in telecom industry: Datasets, methods and metrics." International Research Journal of Engineering and Technology (IRJET) 3.04, 2016.

[2] Xu, Jingxiu, et al. "Early Warning of Telecom Customer Churn Based on Multi-algorithm Model Optimization." Frontiers in Energy Research, 2022, 935.

[3] M. Li, et al. "An early warning model for customer churn prediction in telecommunication sector based on improved bat algorithm to optimize ELM." International Journal of Intelligent Systems 36.7, 2021, 3401 - 3428.

[4] Natekin, Alexey, and Alois Knoll. "Gradient boosting machines, a tutorial." Frontiers in neurorobotics 7, 2013, 21.

[5] Myles, Anthony J., et al. "An introduction to decision tree modeling." Journal of Chemometrics: A Journal of the Chemometrics Society 18.6, 2004, 275 - 285.

[6] Noble, William S. "What is a support vector machine?" Nature biotechnology 24.12, 2006, 1565 - 1567.

[7] Biau, Gérard, and Erwan Scornet. "A random forest guided tour." Test 25, 2016, 197 - 227.

[8] Y. Qiu, et al. "Clustering Analysis for Silent Telecom Customers Based on K-means++." 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC). Vol. 1. IEEE, 2020.

[9] McDonald, Gary C. "Ridge regression." Wiley Interdisciplinary Reviews: Computational Statistics 1.1, 2009, 93 - 100.

[10] Kaggle. Churn in telecoms dataset, 2018. https://www.kaggle.com/becksddf/churn-in-telecoms-dataset/data.