Telcom Churn Prediction

Yashwanth Saladi [1]

Abstract

In the recent decade, there has been tremendous growth in technology and the economy which has led to rapid progress in many industries, especially in the telecom industry where there is growing customers size. Due to the presence of competition, telecom companies provide a wide variety of services to the customers in order to retain their customer base and survive in the current market. It becomes very important for a company to determine which customer will discontinue its service so that the company can use various strategy in order to retain the customer subscription. Collecting the customer data and understanding patterns in it can help in determining customer churn. In this paper, we will discuss various machine learning models such as Logistic Regression, SVM, Random Forest and Gradient boosted tree, etc, which can be trained on the customer churn data and use to make predictions regarding customer churn. We also make comparisons among these models using evaluation metric like F1 score, ROC curve and AUC score.

Keywords — Churn prediction, telecom system, Customer retention, binary classification, supervised learning.

1. Introduction

In the Telecommunication Industry, the revenue is completely based on user subscription. These subscriptions are mostly monthly subscriptions and can be terminated in any month without notice. Due to heavy competition in the market, service availability and offers customers tend to stop subscribing and shift to competitors. Many factors are influencing the churn behaviour of customers. Telecom companies spend a lot of revenue on existing customers to continue subscribing then spending on attracting new customers. So telecom companies take churn prediction as an important task for their business continuity.

There is a lot of research happening in churn prediction over years. The churn prediction is being improved by using data mining techniques and by addressing the issues of handling imbalanced dataset. Recently bagging and boosting techniques are being studied to improve the prediction. And companies are finding ways to improve the accuracy of the predictions. Along with churn prediction, it is also important to identify the factors which are highly influencing the churn rate.

2. Problem statement

The customer churn prediction problem is a binary classification problem. Given the dataset containing features like customer personal details, demographics information, service usage information, etc, we need to identify if the customer will continue using the service from that company or not. When the true label is also provided with the dataset, we can use supervised learning algorithms. We can also find which features actually drive customer churn, which can be used to identify customer ”pain points” and resolve them by providing offers to make customers stay.

3. Proposed work

We propose to solve the problem of churn prediction using machine learning models. We will use python programming to build the models for telecom churn prediction. In our experiment, we plan to build different machine learning models that can analyze the data to predict customer churn. These models are Logistic Regression, Support Vector Machine, Gradient Boosting Trees, simple neural network.

- Dataset:- A telecom dataset is taken from IBM Cognos Analytics 11.1.3+ base samples dataset, dataset contains data form a fictional telco company that provided home phone and Internet services to 7153 customers in California in Q3.data is in .csv format. The dataset contains 7043 observations with 33 variables.
- Data Preparation: The dataset given in csv format is read into dataframe using pandas library.
- Data Preprocessing: Data preprocessing is the most important phase before prediction models as the data may contain some issue like different format of attributes, missing values,etc. These cases need to be handled before training models.

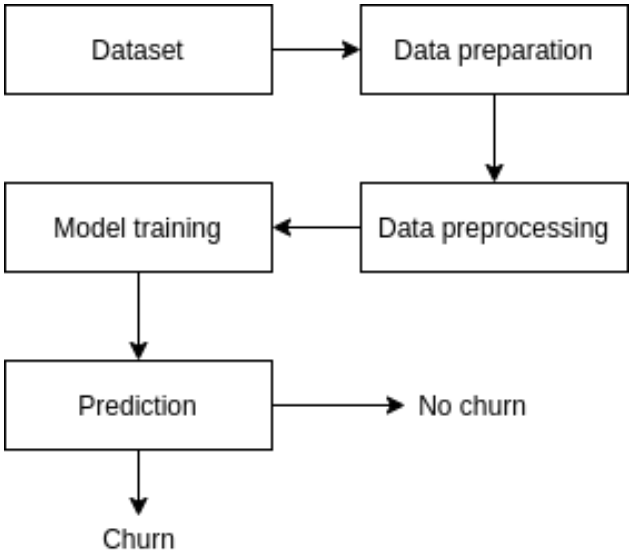


Figure 1. Proposed workflow

- Model training: We train different models on the preprocessed data and compare among the models to pick the best for prediction.

- Decision: Based on test data and classification models we can take a decision whether the customer is churner or not.

4. Intuition

Telecommunications predicting the correct churn behaviour and retaining customers is very expensive. So, misclassification of churn behaviour costs a lot of money. There is a huge explosion of data that can be used by the company to generate valuable insights. Also, the given problems is a binary classification problem. Considering the above statements, machine learning seems to be a suitable approach. Along, with training the models, We need to concentrate on how many of the positive and negative classes are incorrectly classified. Specificity and sensitivity are important for our evaluation which concentrates on false positives and true negatives. We considered using the F1 score, ROC curve and AUC curve as an evaluation metric. F1 score takes into account the precision and recall of a model. Receiving operator characteristics(ROC) curve is a plot that evaluates the performance of the model at all classification thresholds. This graph plots between true positive rate and false-positive rate. The area under the curve(AUC) will be our model evaluation metrics. A model with a good AUC score performs well. For this, we perform different classification algorithms and evaluate AUC-ROC for each model.

5. Experiment

All the experiments were performed on Google Colaboratory notebooks. The given data set contains 7153 observations and 33 attributes. We followed some basic data preprocessing steps. Details about the steps are given below.

5.1. Preprocessing

- We found some of the attributes were not necessary for the problem. These attributes are CustomerID, Count, Country, State, City, Lat Long, Churn Label, Churn Score, CLTV, Churn Reason. We maintained zip code instead of the other location information like country, state, city. We maintained the Lat Long attribute which is the combined latitude and longitude of the customer's primary residence. We maintained churn value which was a direct alternative to churn label. We could you churn the reason by processing it using any natural language processing technique but it was not in the scope of this project. We removed the churn score as it was provided by IBM predictive tool which we didn't want to use.

- We checked for null values in the sample. Samples with null values should be either removed or modified using techniques like averaging, etc. In our data set, we didn't find any samples with null values.

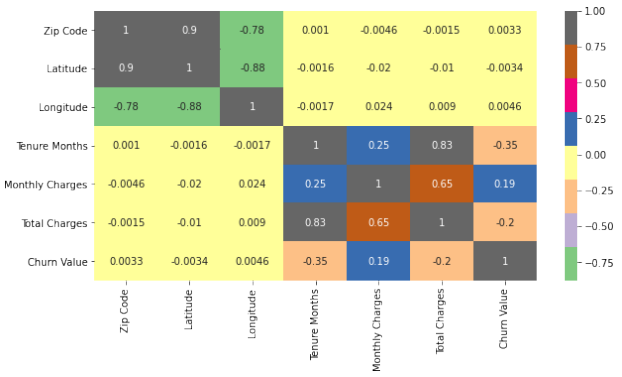


Figure 2. Correlation plot

- We plotted the correlation plot of churn value with respect to other features.

- Most of the attributes are binary/ categorical. We used label encoding in order to convert categorical attributes to binary attributes. Also, there are some features that have values that are more than two-digit. Such values can influence the model performance and to control it we have used normalisation techniques like min-max scaling.

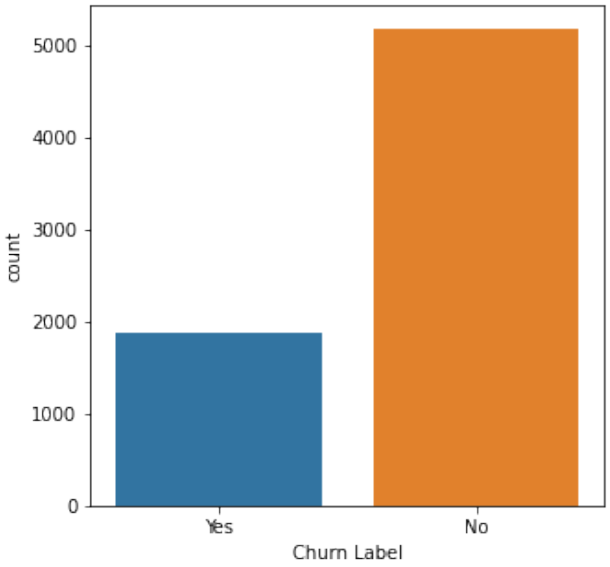


Figure 3. Countplot for churn label

- We found that the data set is imbalanced, so we used Synthetic Minority Oversampling TEchnique, or SMOTE inorder to create a data set with balanced class representation.

5.2. Modelling

The problem of churn prediction is a binary classification problem. The data set contains the features as well as the correct labels. In such case, supervised machine learning algorithms can be used as we can train the models on some samples of data and test the model performance in other samples. We have considered the following supervised learning algorithms

- **K nearest neighbors**
- **Logistic regression**
- **Support vector machine**
- **Decision trees**
- **Random forest**
- **LightGBM**
- **XGBoost**
- **Artificial neural networks**

Table 1. F1 score and AUC score for various models

MODEL	F1 SCORE	AUC SCORE	OPTIMIZED?	RESAMPLED?
KNN1	0.389	0.488	×	×
KNN2	0.315	0.490	√	×
KNN3	0.744	0.811	×	√
KNN4	0.797	0.839	√	√
LOGREG1	0.552	0.650	×	×
LOGREG2	0.552	0.650	√	×
LOGREG3	0.725	0.793	×	√
LOGREG4	0.725	0.793	√	√
DECTREE1	0.528	0.598	×	×
DECTREE2	0.626	0.656	√	×
DECTREE3	0.813	0.859	×	√
DECTREE4	0.837	0.902	√	√
RANFOR1	0.599	0.696	×	×
RANFOR2	0.579	0.672	√	×
RANFOR3	0.874	0.956	×	√
RANFOR4	0.873	0.957	√	√
LGBM1	0.616	0.703	×	×
LGBM2	0.872	0.956	×	√
XGBOOST1	0.627	0.727	×	×
XGBOOST2	0.875	0.958	×	√
ANN	0.607	0.578	×	×

6. Results

We worked on various supervised learning algorithms. For most of the algorithms we trained a model based on default setting provided by sci-kit learn library. We also performed model hyperparameter tuning using GridSearch technique. Inorder to avoid bias due to specific split of the data, we used 5-fold cross validation. As there was imbalance in dataset, we used SMOTE technique inorder to balance the dataset and trained some models on this data. The results from the above combination of models is given in table 1. We have used F1 Score and AUC score as model evalution metrics.

6.1. Receiving operator characteristics

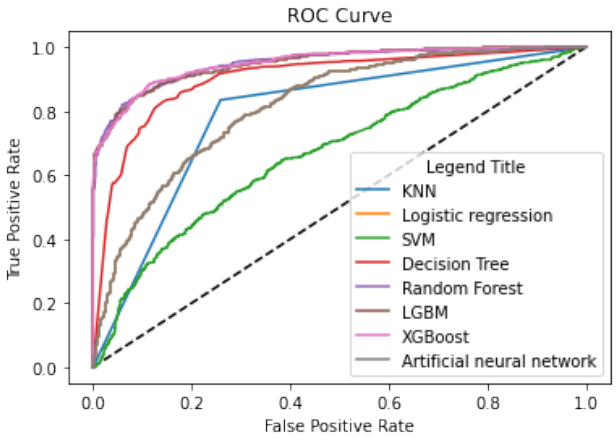
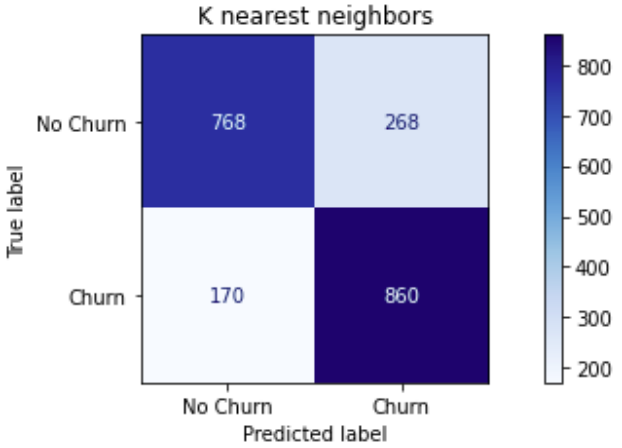
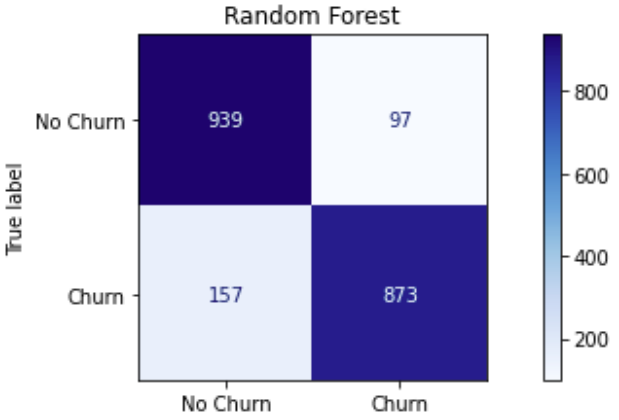


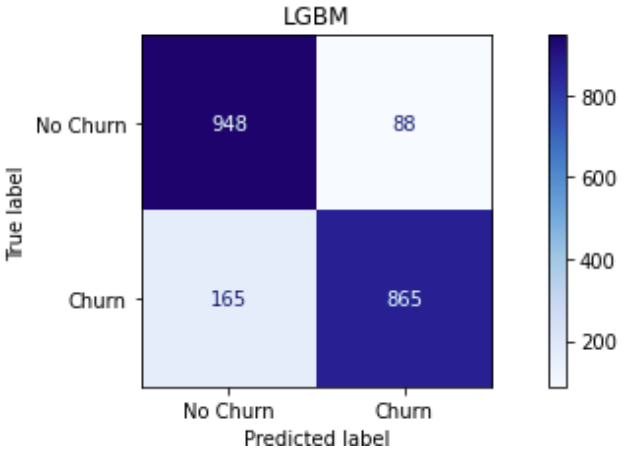
Figure 4. ROC curve

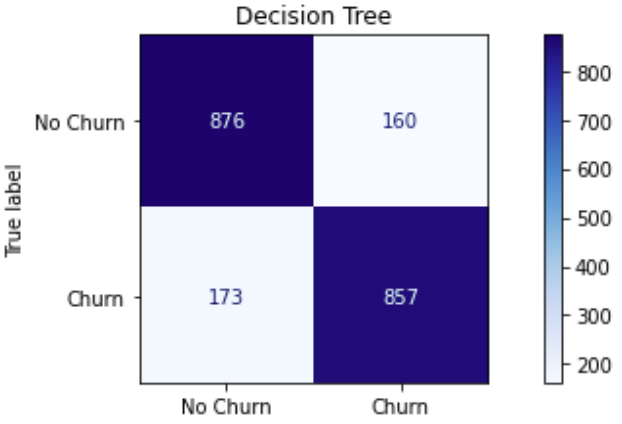
6.2. Observation

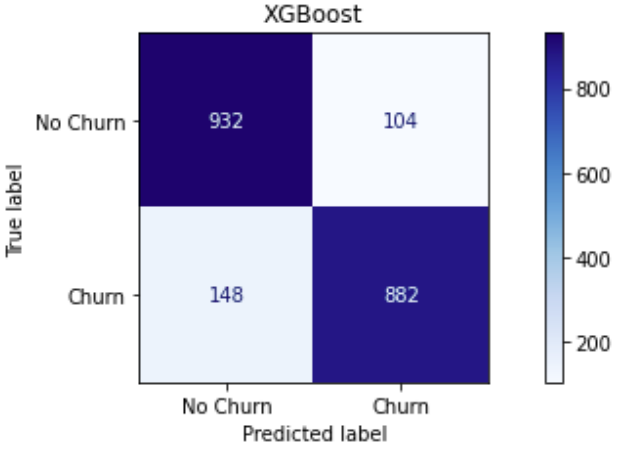
From the above table and plots, we find that models trained on using resampled data have given better F1 score compared to models which where trained on imbalanced data. Overall, the best model is XGBOOST2 which is a model trained using XGBoost algorithm on resampled dataset. It is followed by RANFOR4 which is a model training using random forest algorithm on resampled data and tuned parameters.

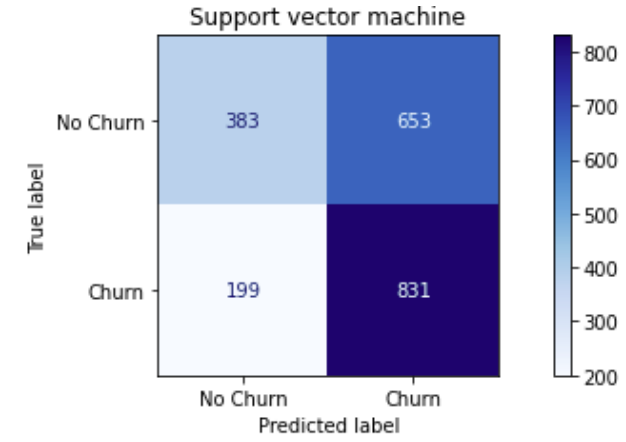


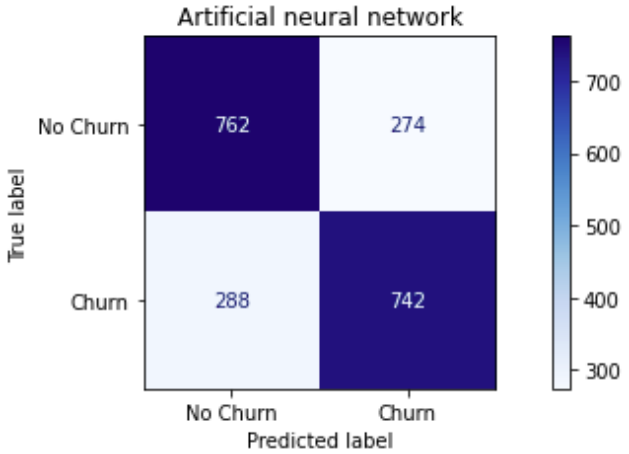












7. Conclusion

There is scope for further improvement in our approach. The dataset is very small which can be one limitation of the machine learning approach. More number of samples can help us in training better models. We could also try different feature engineering techniques in order to create new features which could further help in improving the predictions. We could work with the churn reason feature using natural language processing by training a sentence classification model over the churn reason feature for predicting the churn class. Apart from churn classification, the problem can also be extended to find the importance of each customer in order to roll out customer-specific plans based on the customer importance score.

References

Burez, J., Van den Poel, D.: Handling class imbalance in customer churn prediction. Expert Syst. Appl. 36, 4626–4636 (2009)

De Bock, K.W., Van den Poel, D.: Reconciling performance and interpretability in customer churn prediction using ensemble learning based on generalized additive models. Expert Syst. Appl. 39, 6816–6826 (2012)

Huang, B., Kechadi, M. T., Buckley, B. (2012). Customer churn prediction in telecommunications. Expert Systems with Applications, 39(1), 1414–1425. doi:10.1016/j.eswa.2011.08.024

Keramati, A., Jafari-Marandi, R., Aliannejadi, M., Ahmadian, I., Mozaffari, M., Abbasi, U.: Improved churn prediction in telecommunication industry using data mining techniques. Appl. Softw. Comput. 24, 994–1012 (2014)

A. Gaur and R. Dubey, "Predicting Customer Churn Prediction In Telecom Sector Using Various Machine Learning Techniques," 2018 International Conference on Advanced Computation and Telecommunication (ICACAT), 2018, pp. 1-5, doi: 10.1109/ICACAT.2018.8933783.

K. Dahiya and S. Bhatia, "Customer churn analysis in telecom industry," 2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), 2015, pp. 1-6, doi: 10.1109/ICRITO.2015.7359318.

Ngurah Putu Oka H, Ajib Setyo Arifin, "Telecommunication Service Subscriber Churn Likelihood Prediction Analysis Using Diverse Machine Learning Model", Mechanical Electronics Computer and Industrial Technology (MECnIT) 2020 3rd International Conference on, pp. 24-29, 2020. Ning, L., Hua, L., Jie, L., Guangquan, Z.: A customer churn prediction model in telecom industry using boosting. IEEE Trans. Industr. Inform. 10, 1659–1665 (2012)

Ullah, I., Raza, B., Malik, A. K., Imran, M., Islam, S. ul, Kim, S. W. (2019). A Churn Prediction Model using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector. IEEE Access, 1–1. doi:10.1109/access.2019.2914999