Data Science

Telecom Churn Prediction

Nikhil Mishra; Urjit Patil; Atharva Bhusari; Abhinav Saxena

Abstract

Logistic regression, support vector machine (SVM), and ensemble models were built and evaluated them based on various metrics, including recall, precision, F1 score, and area under the curve (AUC). The primary objective is to obtain the best possible AUC and F1 score while maximizing recall. This would ensure that the selected model has high accuracy in identifying the customers whose churn is positive.

Problem

Predicting churn for a telecom company is crucial for retaining customers and increasing revenue. By identifying customers at risk of leaving and addressing the factors that contribute to churn, such as poor network coverage or billing issues, the company can offer targeted promotions, discounts, or personalized services to improve customer satisfaction and reduce churn. Predicting churn can have a significant impact on the success of a telecom company, making it essential for anyone who cares about the customer experience or the company's success.

Data

Telecom Churn data is obtained from Kaggle, where it contains various columns showcased below:

customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...
7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	...
5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	...
3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	...
7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	...
9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	...

The performed data preprocessing involves the conversion of binary variables into numerical values and the utilization of one-hot encoding to handle categorical variables. The TotalCharges column is converted from a string type to a float type while dropping irrelevant columns. Moreover, standard scaling is applied to numerical columns to ensure that all data is of a comparable scale.

Methodology

In order to predict the churn, we performed EDA to understand the spread of each variable and their collinearity with the target variable. This was followed by building and evaluating a Logistic Regression model, a SVM Classifier model and multiple ensemble classifier models.

The steps followed were:

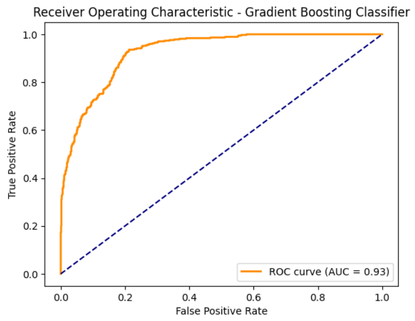
1. Building a vanilla model
2. Finding the optimal threshold for the vanilla model by maximizing the F1 score
3. Performing hyperparameter tuning for the model
4. Finding the optimal threshold for the tuned model by maximizing the F1 score
5. Comparing the models based on Precision, Recall, F1 score, and ROC-AUC score

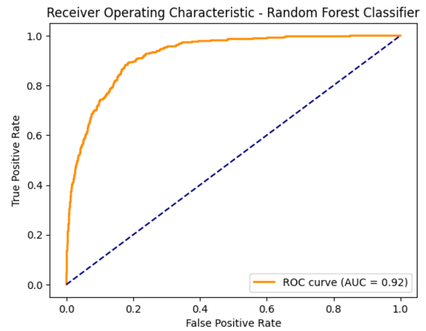
These steps are followed for all the models in order to find the most optimal threshold for each algorithm. By doing this, we ensure that the probability on which the model is deciding the target variable is not set as 0.5, instead the optimal probability is used to decide the class.

Post this we also tried implementing voting classifier to get the prediction of the best model from all the models built and tried to implement a metric for selecting that as well.

Findings and Evaluation

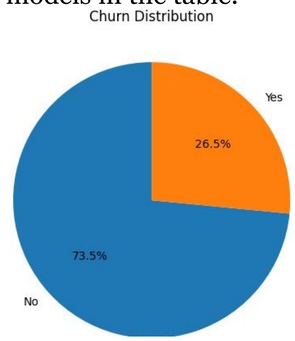
In the first table, the top performing models in terms of F1 score are the Tuned Threshold Random Forest Classifier (0.744) and the Tuned Threshold Gradient Boosting Classifier (0.742). These models also have high recall scores (0.880 and 0.920 respectively), indicating that they are effective at identifying positive instances. However, their precision scores (0.645 and 0.622 respectively) are somewhat lower, suggesting that they may produce a relatively high number of false positives.

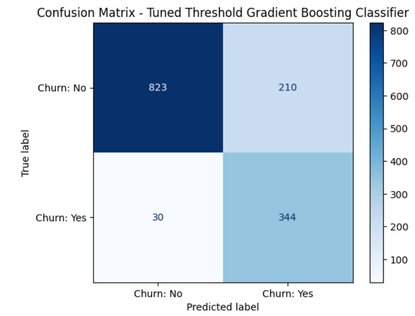




In the second table, the top performing models in terms of F1 score are the Tuned Threshold SVM Classifier (0.744) and the Optimal Threshold SVM Classifier (0.742). These models also have high recall scores (0.802 and 0.805 respectively), but their precision scores (0.694 and 0.689 respectively) are somewhat lower.

The Optimal Threshold Logistic Regression model has the highest precision score (0.703) of all the models, but its F1 score (0.726) is somewhat lower than the top performing models in the table.





Model	Recall	Precision	F1 Score	ROC-AUC Score
Tuned Threshold SVM Classifier	0.794118	0.693925	0.740648	0.833651
Optimal Threshold SVM Classifier	0.799465	0.688940	0.740099	0.834389
Optimal Threshold Logistic Regression	0.751337	0.699005	0.724227	0.817101
Tuned Threshold Logistic Regression	0.828877	0.637860	0.720930	0.829250
Logistic Regression	0.711230	0.722826	0.716981	0.806244
Tuned Logistic Regression	0.700535	0.715847	0.708108	0.799929
Tuned SVM Classifier	0.681818	0.732759	0.706371	0.795895
SVM Classifier	0.671123	0.733918	0.701117	0.791515

Conclusion

After building and evaluating the variations of models, the two best models were the Tuned Threshold Random Forest Classifier and the Tuned Threshold Gradient Boosting Classifier. Among the top 2 models, the Tuned Threshold Gradient Boosting model will be selected as the best model as it provides a better Recall than the Tuned Threshold Random Forest Classifier model. This edge in the Recall is obtained with a slight drop in Precision, which is acceptable in this particular use case where the primary aim is to predict the customers for whom the Churn is positive. A superior AUC score for the Tuned Threshold Gradient Boosting Classifier model further bolsters the claim of it being better than the Tuned Threshold Random Forest Classifier.