Telecom Churn Case Study

Building a predictive model to identify churn in telecom industry.

SRUJITHA REDDY MEKALA
G01326037, George Mason University, smekala2@gmu.edu

MANANKUMAR THAKKAR
G01340764, George Mason University, mthakkar@gmu.edu

SUDEEP STEPHEN YALLA
G01338558, George Mason University, syalla@gmu.edu

Abstract

This project aims to analyze the customer-level data of a major telecom provider, build predictive models to identify customers at risk of churn and identify the factors associated with churn. We also try to achieve good accuracy without compromising the sensitivity. Here we are using the usage-based definition to define churn. Three data mining techniques, logistic regression, decision tree, and Support Vector Machine are used to perform the analysis. Initially, the data is pre-processed and then Exploratory Data Analysis is performed to elicit valuable insights. The churn prediction percentage appeared to be around 3% which indicated a class-imbalance. This is nullified using the SMOTE technique. Then, new features are derived, and the above-mentioned models are trained along with tuning the hyper-parameters. Finally, appropriate evaluation metrics are used to evaluate the models. The decision tree classifier provided both accuracy and sensitivity compared to logistic regression and SVM.

Additional Keywords and Phrases: Churn prediction, Logistic Regression, Decision Tree, Support Vector Machine, Classification.

1 Introduction

Predicting churn is basically identifying the customer who are about to leave a particular service or a service provider. The telecom business has churn rates of 15-25% as a result of intense competition. This has been a major problem in the telecom industry as the cost of acquiring a new customer is 5-10 times higher than retaining an already existing customer. Hence, predicting the customers who are at the risk of churn and taking necessary steps would help the service providers.

The dataset used in this project is the telecom churn data from Kaggle. This dataset has 99999 rows indicating the individual users and 226 columns indicating different attributes. The attributes include detail like mobile number, circle ID, account information like billing cycles, payment details etc and demographic information like age, gender etc. The dataset has the user-level information for 4 months, June through September, which are encoded as 6,7,8 and 9 respectively.

Based on this dataset, we are using the usage-based approach to define churn. Usage-based churn implies analyzing user data including who have not made any calls or accesses the internet over a particular period. The goal is to determine churn in the last month based on the data available for the first 3 months.

The data is preprocessed i.e., missing data is imputed manually and columns with more than 70% missing data are dropped. Columns containing dates, area and unique values are eliminated since they do not help in the analysis. Further high-value customers are filtered, and attributes of the churning phase are eliminated. Finally, feature selection is done using RFE technique. This pre-processed data is used to train logistic regression classifier and perform the analysis. In case of decision trees, k-fold method is employed to split the training data and the best possible combination of attributes are found to perform the analysis along with cross-validation. The "svm" function from the sklearn library is used to analyze the pre-processed data and classify using Support Vector Machine. Although all the 3 models showed significant accuracy which is around 90%, the decision tree provided higher sensitivity than logistic regression and support vector machine.

2 METHOD

Predictive modeling is primarily concerned with forecasting how a client will act in the future based on their prevision behaviour [1]. Customer Relationship Management (CRM) data and DM are analyzed using the predictive modelling to create customer-level models that indicate the possibility that a customer would perform a specific action. The actions are often connected to sales, marketing, and client retention [1].

The proposed model comprises four steps. Identifying the problem, data selection, data pre-processing, classification. Data pre-processing involves handling missing data, filtering the unwanted data like columns containing dates or unique values etc. It also includes handling the class imbalance condition using SMOTE technique. Here traditional techniques are used to perform the classification.

2.1 Traditional Techniques

2.1.1 **Logistic Regression:** It is used to estimate probabilities and explain the connection between the dependent variable and one or more independent variables. It is used to assist in predicting the chances of occurrence of an event or a decision making. Here it is used to predict the customer churn based on few highly correlated attributes with the target variable. We have used the "LogisticRegression" function from the "sklearn" library. Based on different probability cut offs for churn, separate columns are created, and accuracy, sensitivity and specificity are calculated. The plot of these parameters with respect to probability provided an optimal probability i.e., 0.5 where all the three parameters are balanced.

2.1.2 **Decision Tree:** Decision tree can be used for both classification as well as regression following a tree structure. The predictions are generated by a series of attribute-based splits. Generally, the training data is first partitioned based on the features and then noisy data is removed [2]. Here, the "DecisionTreeClassifier" function is imported from the sklearn library. K-Folds cross validator is used to split the data into train and test sets. Here the train data is split into 5 folds. "GridSerachCV" is used to consider all the possible parameter combinations and choose the best ones. When fitting it on a data set all the possible combinations of parameter values are evaluated and the best combination is retained. It generates candidates from a grid of parameter values specified with the param_grid parameter. To perform cross validation the "cross_val_score" function is used.

2.1.3 **Support Vector Machine:** This is also a machine learning algorithm which can be used for both classification and regression. It finds a hyperplane that classifies the given set of data points with maximum margin or distance

between the plane and the data points. The data points that lie on either side of the plane belong to different classes. Here we use the "svm" function to perform the analysis [3].

3 RESULTS

The graph here shows a 3% churn which is an indicator of class imbalance. If this data is not processed, better accuracy can be obtained by predicting the majority class correctly i.e., no churn class, but sensitivity cannot be obtained since churn class may not be correctly identified and predicted.

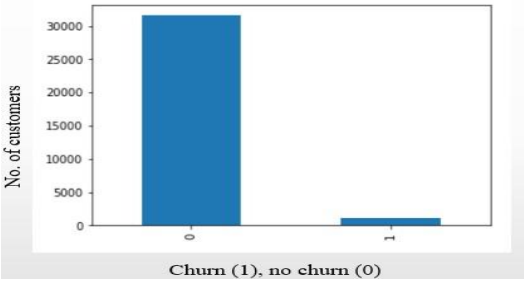

Figure 1: Plot indicating class imbalance.

The graph here shows a plot of all the metrics vs probability of churn. This is used to pick a threshold probability to achieve good accuracy as well as sensitivity in case of logistic regression. Here the threshold or optimal probability is chosen as 0.5.

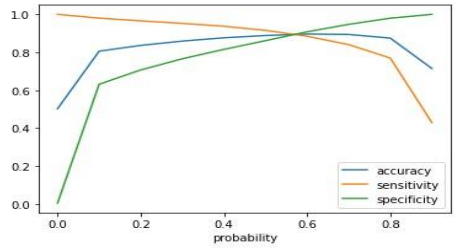

Figure 2: Probability vs Metrics graph

An ROC curve for the logistic regression model with a threshold probability of 0.5 provided an area of 0.96 which indicates good performance or good measure of separability between the classes.

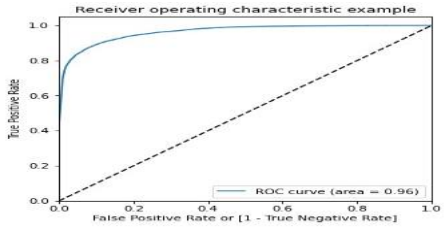

Figure 3: Receiver Operating Characteristics Curve

The following image shows the summary from the logistic regression classifier for the features selected using RFE.

Generalized Linear Model Regression Results

Dep. Variable:	churn	**No. Observations:**	50498
Model:	GLM	**Df Residuals:**	50477
Model Family:	Binomial	**Df Model:**	20
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	nan
Date:	Thu, 12 May 2022	**Deviance:**	nan
Time:	15:27:05	**Pearson chi2:**	4.50e+15
No. Iterations:	100		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.23e+04	1.06e+05	0.116	0.908	-1.96e+05	2.2e+05
std_og_t2t_mou_8	0.2198	0.027	8.111	0.000	0.167	0.273
isd_og_mou_8	-2.8053	0.449	-6.252	0.000	-3.685	-1.926
og_others_8	-2059.4570	2.96e+05	-0.007	0.994	-5.83e+05	5.79e+05
total_og_mou_8	-0.9392	0.034	-27.724	0.000	-1.006	-0.873
loc_ic_mou_8	-3.4291	0.061	-56.091	0.000	-3.549	-3.309
total_rech_data_6	9.204e+04	2.16e+06	0.043	0.966	-4.14e+06	4.33e+06
total_rech_data_7	3.065e+04	2.54e+06	0.012	0.990	-4.95e+06	5.02e+06
total_rech_data_8	1.981e+05	2.99e+06	0.066	0.947	-5.65e+06	6.05e+06
monthly_2g_6	-1.487e+04	3.49e+05	-0.043	0.966	-6.99e+05	6.69e+05
monthly_2g_7	-4552.5172	3.78e+05	-0.012	0.990	-7.45e+05	7.36e+05
monthly_2g_8	-2.848e+04	4.29e+05	-0.066	0.947	-8.69e+05	8.12e+05
sachet_2g_6	-7.082e+04	1.66e+06	-0.043	0.966	-3.33e+06	3.19e+06
sachet_2g_7	-2.342e+04	1.94e+06	-0.012	0.990	-3.83e+06	3.79e+06
sachet_2g_8	-1.502e+05	2.26e+06	-0.066	0.947	-4.59e+06	4.29e+06
monthly_3g_6	-2.145e+04	5.04e+05	-0.043	0.966	-1.01e+06	9.66e+05
monthly_3g_7	-7186.7197	5.96e+05	-0.012	0.990	-1.18e+06	1.16e+06
monthly_3g_8	-4.432e+04	6.68e+05	-0.066	0.947	-1.35e+06	1.26e+06
sachet_3g_6	-3.451e+04	8.1e+05	-0.043	0.966	-1.62e+06	1.55e+06
sachet_3g_7	-1.149e+04	9.54e+05	-0.012	0.990	-1.88e+06	1.86e+06
sachet_3g_8	-7.481e+04	1.13e+06	-0.066	0.947	-2.28e+06	2.13e+06

Figure 4: Summary from logistic regression model.

Table 1: Metrics obtained with different models

Model\Metrics	Data	Accuracy	Sensitivity	Specificity
Logistic Regression	Train data	0.8962	0.8845	0.9078
	Test data	0.8984	0.5529	0.9109
Decision Tree	Train data	0.9485	0.9704	0.9266
	Test data	0.9217	0.8157	0.9255
Support Vector Machine	Train data	0.9533	0.9744	0.9322
	Test data	0.9223	0.6929	0.9306

4 CONCLUSION

In this project a simple churn prediction model is developed using three data mining techniques, logistic regression, decision tree, and SVM. The data set used has user-level information for four months with 99,999 instances and 226 attributes. The metrics evaluation indicates that the decision tree model is the best in hand in terms of accuracy and sensitivity. Features that are highly correlated with the target variable are found using different techniques like RFE, GridSearchCV. Distribution of class is not homogeneous with the other class, which is known as “class imbalance” is handled. Other soft-computing techniques like neural networks or bayes classifier can also be used to build the model. Further, boosting algorithms can be applied to make the model much more efficient.

REFERENCES

[1] Essam Shaaban, Yehia Helmy, Ayman Khedr, and Mona Nasr. July 2012. A proposed churn prediction model, International Journal of Engineering Research and Applications.

[2] Anshul Saini. August 29,2021. Decision tree algorithm – A complete guide.
https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/

[3] Rohith Gandhi. June 7,2018. Support vector Machine- Introduction to machine learning algorithms.
https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

Video Presentation link: https://drive.google.com/file/d/1WT_M6NRorff1n3y8bL1fhl3q4ISac9ZC/view?usp=sharing