Customer Churn Prediction and Classification

Priya Jani
Ahmedabad University
Ahmedabad, India
priya.j@ahduni.edu.in

Yansi Memdani
Ahmedabad University
Ahmedabad, India
yansi.m@ahduni.edu.in

Priyanshu Pathak
Ahmedabad University
Ahmedabad, India
priyanshu.p@ahduni.edu.in

Mohnish Mirchandani
Ahmedabad University
Ahmedabad, India
mohnish.m@ahduni.edu.in

Abstract – This report presents a study on customer churn prediction using machine learning. The objective is to test out various classical machine learning algorithms present in order to predict customer churn accurately. The scope of this report extends to further data analysis and principal component analysis. It also tries to exhaustively compare algorithms and the effects of data refining on similar algorithms.

Keywords - Customer Churn, Classification, Prediction, Logistic Regression, Support Vector Machine, Naive Bayes Classification, Random Forest Classification, SMOTE Analysis, Principal Component Analysis

I. INTRODUCTION

Customer churn, also referred to as subscriber churn or logo churn, refers to the proportion of subscribers who terminate their subscriptions and is commonly expressed as a percentage. Customer churn prediction and analysis is one of the foremost and widespread applications of classical machine learning. Customer churn is a critical metric that can display customer satisfaction at the macro scale. Additionally, the telecom sector generally sees more significant churn rates than other sectors. This creates a large-scale requirement for better prediction models.

II. LITERATURE SURVEY

Customer Churn prediction has been a problem of the data transfer age with prominent research across the problem statement. Network analysis for customer churn is one-way networks connected by the similarity of churning members used to predict the possibility of churn. Classical Machine Learning algorithms provide a proven approach to predicting churn. Deep Learning models tend to go one step further to predict churn rates.

III. IMPLEMENTATION

For the purpose of training the model, the following was implemented in sequential order:

A. Data cleaning: On checking for duplicate and missing values, we found the data accurate and consistent.

B. Exploratory Data Analysis and Data Preprocessing: Conversion of categorical features to numerical features. Trend analysis of each feature with churn rate (y). Data unit conversion where required.

C. Correlation: Correlation matrix to find linear relationships between two variables.

D. Generalized Linear Model: Relations between predictor variables and response variables devised based on the p-values.

E. Feature Scaling: Used to standardise the independent features within a fixed range.

F. Classification Models

For the three models we have used, the approaches are as follows:

1) *Binary Logistic Regression*
2) *Support Vector Machine (SVM)*
3) *Naive Bayes Classifier*
4) *Random Forest Classification*

G. Smote Data Balancing: Used to balance imbalance dataset.

H. Correlation based Feature Selection: Used to select features based on their impact and similarity.

I. Principal Component Analysis: Used for dimensionality reduction and feature selection

J. Logistic Regression from scratch: To understand the default function's inbuild implementation.

K. SVC from scratch: To understand the default function's inbuild implementation and changing the parameters.

L. Naive Bayes from scratch: To understand the default function's inbuild implementation.

M. ROC-AUC Curve: Evaluation metric for models

IV. RESULTS

The models are evaluated based on the following factors:

Before Data Balancing:

	Logistic Regression (%)	Support Vector Classifier (%)	Naive Bayes Classifier (%)	Random Forest Classifier (%)
Accuracy	81.27	80.52	65.26	79.71
Precision	66.04	66.42	41.91	64.03
Recall	57.74	50.81	86.88	50.27
F1 Score	80.85	79.61	67.28	78.86

Table1 Before Data Balancing

After Data Balancing:

	Logistic Regression (%)	Support Vector Classifier (%)	Naive Bayes Classifier (%)	Random Forest Classifier (%)
Accuracy	82.05	82.25	67.89	83.52
Precision	64.07	66.66	39.63	68.02
Recall	46.21	42.42	81.36	50.60
F1 Score	80.94	80.73	70.53	82.60

Table2 After Data Balancing

After Feature Selection based on Correlation:

	Logistic Regression (%)	Support Vector Classifier (%)	Naive Bayes Classifier (%)	Random Forest Classifier (%)
Accuracy	80.90	80.47	73.55	78.67
Precision	65.40	66.99	49.47	61.48
Recall	56.46	49.18	76.86	48.26
F1 Score	80.42	79.41	74.99	77.77

Table2 Features selected using correlation matrix

After PCA:

	Logistic Regression (%)	Support Vector Classifier (%)	Naive Bayes Classifier (%)	Random Forest Classifier (%)
Accuracy	80.07	80.10	79.66	81.30
Precision	59.40	63.50	55.63	61.52
Recall	36.36	27.42	47.87	45.30
F1 Score	78.20	76.92	79.10	80.22

Tab4 After PCA

ROC-AUC Curves:

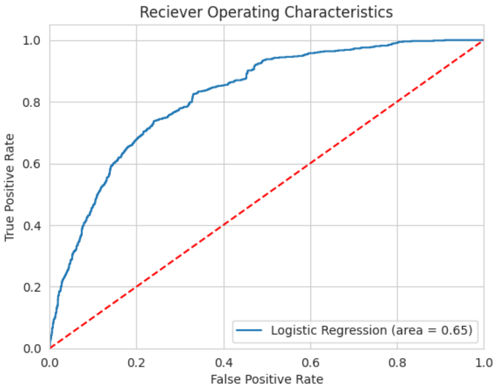

Fig1 Logistic Regression

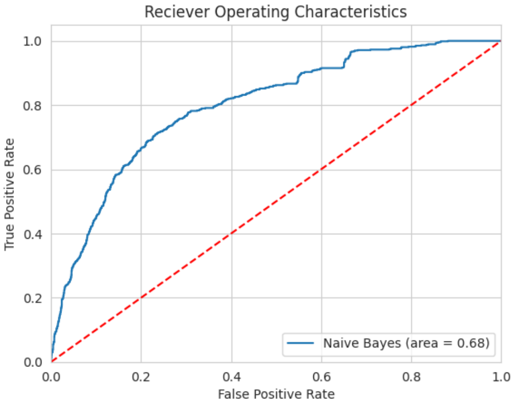

Fig2 Naive Bayes

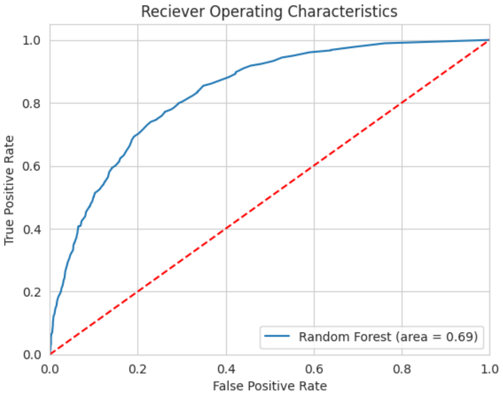

Fig3 Random Forest

The comparison on the basis of training time and training accuracy when changed some parameters and specifications on logistic model is given below.

LR model	training time (9769 records)	training accuracy
Loss function + Gradient descent	61.11 seconds	73.20% (threshold=0.8)
MLE + Gradient ascent	60.61 seconds	5.47%
sklearn	124.14 seconds	66.15%

Table5 LR model Analysis

The comparison when kernel was changed on the SVC model is given below:

Evaluation: Polynomial kernel (Accuracy: 0.77 (Support: 1954))

Label	precision	recall	f1-score	support
0	0.77	1.00	0.87	1513
1	0.00	0.00	0.00	441

Table6.1 Polynomial kernel

Evaluation: RBF kernel (Accuracy= 0.82 (Support: 1954))

Label	precision	recall	f1-score	support
0	0.85	0.94	0.89	1513
1	0.66	0.43	0.52	441

Table6.2 RBF kernel

Evaluation: Sigmoid kernel (Accuracy= 0.82 (Support: 1954))

Label	precision	recall	f1-score	support
0	0.86	0.92	0.89	1513
1	0.65	0.50	0.56	441

Table6.3 Sigmoid kernel

Evaluation: Linear kernel (Accuracy= 0.83 (Support: 1954))

Label	precision	recall	f1-score	support
0	0.86	0.92	0.89	1513
1	0.65	0.49	0.56	441

Table6.4 Linear kernel

Grid-search is used here to identify a model's ideal hyperparameters, or those that produce the most "correct" predictions. After Grid-Search, the classification report of SVC is as below.

Accuracy= 0.82 (Support: 1954)

Label	precision	recall	f1-score	support
0	0.85	0.94	0.89	1513
1	0.67	0.43	0.52	441

Table6.5 SVC Classification Report

V. CONCLUSION

We observed that the data balancing technique that is “smote analysis” was quite effective in our case as we had imbalance in the churn data. The results after data balancing were good.

We observed varied results across the algorithms. Each has its own merits and demerits. Wether using correlation or PCA, there is a straight increasing trend seen in the accuracy results of Naive Bayes, and a straight decreasing trend is seen in Logistic Regression.
After PCA, The final values of accuracy, f1 score, and precision had less impact. However, recall values are the most affected, which states that the machine learning models slightly failed to create annotations that it should have created.

The logistic model was analysed on the basis of training time and training accuracy. MLE+ gradient ascent takes the least amount of time but failed to give good accuracy. However, Loss function + gradient descent is much better than the rest. It was observed that as the threshold was increased, the training accuracy also increased.
The parameters of SVC model were changed and observed. The accuracy was best when a linear kernel was used. On-Grid Search, the parameters for the best estimator were: C=1 and gamma=0.1, where C is the regularization parameter and gamma controls the shape of the kernel function.

The models were analysed using ROC-AUC curve. AUC stands for the level or measurement of separability, and ROC is a probability curve. It reveals how well the model can differentiate across classes. It was observed that the AUC value of random forest was maximum (AUC=0.69). This was also observed from the table

VI. REFERENCES

[1]“12.1 - Logistic Regression | STAT 462,” *12.1 - Logistic Regression | STAT 462*. [Online]. Available: https://online.stat.psu.edu/stat462/node/207/

[2]S. R. Publishing, “Churn Prediction Using Machine Learning and Recommendations Plans for Telecoms,” *Churn Prediction Using Machine Learning and Recommendations Plans for Telecoms*, Nov. 05, 2019. [Online]. Available: https://www.scirp.org/html/3-1731142_96177.htm

[3]“Naive Bayes Classifier in Machine Learning - Javatpoint,” *www.javatpoint.com*. [Online]. Available: https://www.javatpoint.com/machine-learning-naive-bayes-classifier

[4]Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019, March 20). Customer churn prediction in telecom using machine learning in big data platform. Journal of Big Data, 6(1). https://doi.org/10.1186/s40537-019-0191-6

[5]*sklearn.model_selection.GridSearchCV*. (n.d.). Scikit-learn. https://scikit-learn/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

[6]Hajian-Tilaki, K. (n.d.). Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. PubMed Central (PMC). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3755824/

[7]*What is Random Forest? | IBM*. (n.d.). What Is Random Forest? | IBM. https://www.ibm.com/in-en/topics/random-forest