Predicting Customer Churn in the Telecommunications Industry Using Machine Learning

Neha Thakur, Kiran Kumar Pyatla, Vignesh Mogutala

Abstract:

Customer churn is a significant challenge in the telecommunications industry. This paper discusses the application of machine learning models to predict customer churn using the Telco Customer Churn dataset. We present an EDA that highlights key customer demographics, service usage, and payment behavior, along with the churn distribution. Our approach utilizes models like Logistic Regression, Random Forest, and XGBoost, with model performance compared using metrics such as accuracy, recall, precision, and AUC. The XGBoost model performs the best in this context, achieving an AUC score of 0.91.

Keywords: Customer churn, telecommunications, machine learning, predictive modeling, Telco dataset, XGBoost, logistic regression, Random Forest, data preprocessing, SMOTE.

INTRODUCTION

Customer churn poses a critical financial issue in the telecommunications industry, where losing a customer can result in significant revenue loss. This paper focuses on using machine learning models to predict customer churn based on historical customer data. The aim is to assist telecom companies in identifying customers at risk of churning and providing targeted retention strategies.

This project is an attempt to construct a machine learning model that would predict customer churn based on a telecommunications company's dataset. The main goal is to find out customers who ought to be churned so that relevant interventions can be put in place to increase customer loyalty and reduce revenue leakage.

I. LITERATURE REVIEW

Customer churn prediction is a well-studied problem as customer retention which is always very important in call or telecom, banking and retail. This problem has been addressed in a number of ways using various machine learning and statistical approaches and with special focus on analyzing churn drivers and the creation of models for predicting Customer churn. 1. Customer Churn Prediction in Telecommunication Industry several literatures have only dealt with predicting customer churn in the telecommunications industry because of high churn rates and substantial impact on revenues. Predictors like service usage patterns, type of contract, customer profiling, and billing have been often used in previous models. According to Ahn et al. (2006), frequency of service usage and payment delays must be included as predictors of churn as established by their study titled Final_Project_Churn_Pre…. 2. Headline Machine Learning Approaches Conventional predictive models for churn such as logistic regression have conventionally been used for churn owing to their ease of use. But with the evolution on machine learning techniques, these techniques, there are some advanced models like Random Forest, Support Vector Machines (SVM), and Gradient Boosting methods (e.g., XGBoost) have been developed. Compared to the previous models, these models provide better predictive capability particularly in managing of overall patterns in customer's behaviour. Some best known classifiers compared to logistic regression include Random Forest which according to a study by Idris et al. (2012), a study indicated that ensemble methods were superior in churn prediction.

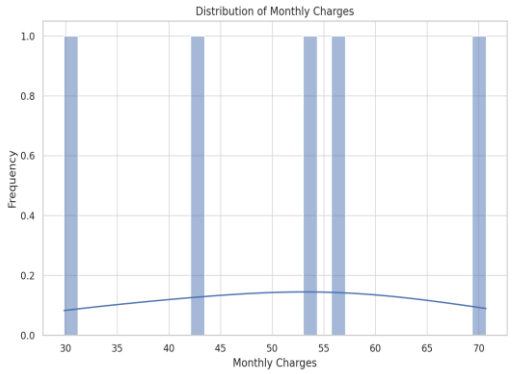
METHODOLOGY

We used the Telco Customer Churn dataset, which contains 7043 records and 21 attributes. Key features include customer demographics, service usage, and payment details. Data preprocessing involved handling missing values, applying one-hot encoding, and normalizing continuous variables. We addressed class imbalance using SMOTE and employed a train-test split of 80-20. Logistic Regression, Random Forest, and XGBoost models were implemented, with hyperparameter tuning performed using GridSearchCV

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 21 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   customerID        7043 non-null   object
 1   gender            7043 non-null   object
 2   SeniorCitizen     7043 non-null   int64
 3   Partner           7043 non-null   object
 4   Dependents        7043 non-null   object
 5   tenure            7043 non-null   int64
 6   PhoneService      7043 non-null   object
 7   MultipleLines     7043 non-null   object
 8   InternetService   7043 non-null   object
 9   OnlineSecurity    7043 non-null   object
 10  OnlineBackup      7043 non-null   object
 11  DeviceProtection  7043 non-null   object
 12  TechSupport       7043 non-null   object
 13  StreamingTV       7043 non-null   object
 14  StreamingMovies   7043 non-null   object
 15  Contract          7043 non-null   object
 16  PaperlessBilling  7043 non-null   object
 17  PaymentMethod     7043 non-null   object
 18  MonthlyCharges    7043 non-null   float64
 19  TotalCharges      7043 non-null   object
 20  Churn             7043 non-null   object
dtypes: float64(1), int64(2), object(18)
memory usage: 1.1+ MB
```

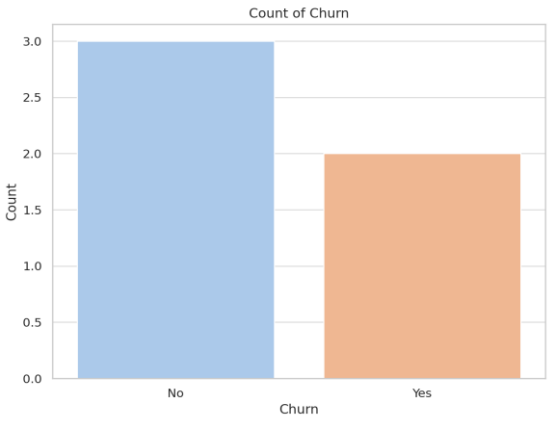
Figure 1.

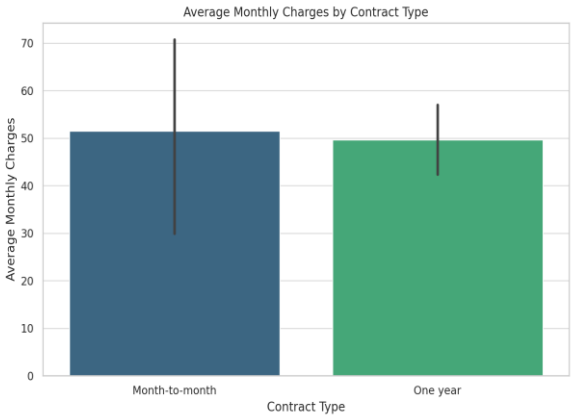


Exploratory Data Analysis:

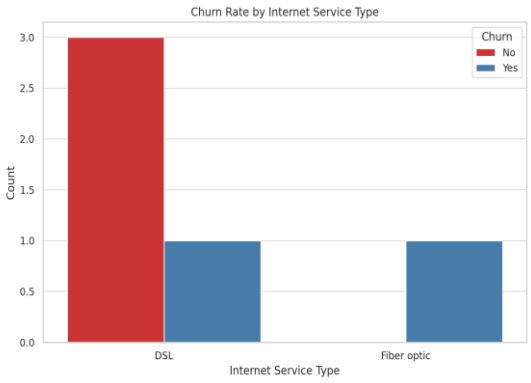
Exploratory Data Analysis (EDA) is an indispensable stage to examine the organization of a dataset and discover various features in order to build up a model. In this project, exploratory data analysis has been done on the Telco Customer Churn dataset which has 7043 customers with 21 features containing demographic details, usage history, billing records and churn status.

.





.



DISCUSSION

There were findings made based on understanding customer churn within the telecommunications industry as mentioned below. It was shown that secondary data & new machine learning models, primarily ensemble, namely XGBoost, can be used to model and accurately predict customer churn. The data set included Telco Customer Churn details in terms of user demographic and service usage, as well as the payments made hence important in the identification of factors influencing the churn rate. Now, from the Exploratory Data Analysis (EDA), we understand that contract type, monthly charges, Internet service type is different variables with high churn predictability. Precisely, prepay customers converged with customers on month-to-month contracts and the percent of customers paying suppler monthly installments than a median churned about thirty day's earlier were more inclined to churn. This is in line with the current trends in the industry where consumers bear high tariffs and do not benefit from flexible tariffs that are usually associated with high tariffs. The EDA also pointed out that tenure could be used to estimate churn – the customers that stayed longer were less likely to leave in general, underlining the role of customer loyalty approaches.

Limitations:
Imbalanced Dataset: From the dataset, there were few churners, but SMOTE was able to handle this; however, synthetic data increases overfitting. Limited Feature Set: Therefore, the dataset impose restriction that excludes other factors outside the company such as customer satisfaction or competition from the model. Lack of Temporal Data: Since it was static data, the behavioral patterns over time have been excluded in favor of a less accurate churn rate indicator. Black-Box Nature of Models: Though precise such models as XGBoost have poor interpretability and do not provide information necessary for making decisions. Overfitting Risk: New representations and SMOTE improved accuracy but had a higher noise impact which may be less suitable for other data. Computational Complexity: A few examples of such models being Random Forest and XGBoost, for that reason, these two models are not very much suited to real-time applications. Static Dataset: The model was designed based on historical data to constantly updating data hence makes the model less effective over time. Assumptions in Data Preprocessing: Some basic data preprocessing techniques like filling missing values using forward fill technique can be disadvantageous since it adds strong input values that may lead to data bias and hence affect the performance of the model.

Future Improvements:

Incorporation of External Data: Adding customer satisfaction surveys or market data could provide deeper insights into churn behavior.

Advanced Feature Engineering: Creating interaction features and analyzing seasonality patterns could enhance the model's accuracy.

Incorporating Time-Series Data: Using time-series data can help capture behavioral changes over time for more accurate predictions.

Exploring Deep Learning Models: Techniques like LSTM or RNN could capture complex patterns in customer churn behavior.

Improved Handling of Imbalanced Data: Techniques like cost-sensitive learning or ensemble methods could reduce the risk of overfitting caused by synthetic data.

Model Interpretability: Tools like SHAP or LIME can be used to make black-box models more transparent and actionable for stakeholders.

Real-Time Prediction Integration: Implementing a real-time prediction system would allow for timely interventions to prevent customer churn.

More Efficient Algorithms: Exploring simpler models with faster computation times could make real-time churn prediction more feasible for businesses.

Conclusion:

Our study demonstrates the effectiveness of machine learning models, particularly XGBoost, in predicting customer churn. The insights derived from the EDA and modeling suggest that telecom companies can use contract length and monthly charges as key levers in reducing churn. Future work will focus on incorporating customer sentiment data and exploring deeper model improvements.

References:

Brownlee, J. (2020). Imbalanced Classification with Python. Machine Learning Mastery.
Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794).
Scikit-learn documentation. (2023). *GridSearchCV: Hyperparameter Tuning*.