Telecom Churn Prediction

Ayush Nilesh Sojitra
Computer Science
The University of Texas at Dallas
Richardson, Texas
AXS220055

Bhavan Girishchandra modha
Computer Science
The University of Texas at Dallas
Richardson, Texas
BGM210003

Darsh Maheshkumar Vaghasia
Computer Science
The University of Texas at Dallas
Richardson, Texas
DXV210035

***Abstract*— The telecommunications industry faces the constant challenge of customer churn, requiring proactive strategies for retention. This paper introduces an innovative approach to telecom churn prediction, leveraging PySpark's distributed computing capabilities and implementing logistic regression with stochastic gradient descent (SGD). The study encompasses a comprehensive preprocessing pipeline, including one-hot encoding and feature scaling. Logistic regression is chosen for its suitability in binary classification tasks, while SGD ensures scalability in large-scale datasets. The model achieves a commendable 75% accuracy on test data, showcasing the efficacy of the proposed methodology. The findings contribute to the field by providing telecom providers with a scalable and efficient tool for customer churn prediction, enabling targeted retention efforts and minimizing revenue loss.**

Keywords*— *PySpark, Logistic Regression, Scalability, Stochastic Gradient Descent, Customer Behavior, Telecom Churn

I. Introduction

The industry of telecommunication is extremely competitive with each other. Providers are trying to retain the customers by giving different offers and customer satisfaction all better than other providers. There are big offers provided by the provider if the person switches the provider. To cope up in this fierce competition Churn Prediction has played a great role getting to identify the customers which are most likely to leave or discontinue the service by following some sort of trend by observing various aspects. This prediction of whether the customer will continue or leave and getting insights like the quarter where the provider usually loses the maximum number of customers, help the providers to come up with some retention strategy to hold on the customers. As we consider more and more aspects which are basically the features of our model, the computation of the training of model gets more complex. In this paper we are analyzing different aspects which influence the retention of a customer and build a logistic regression churn prediction model using spark. With the help of pyspark which is based on Hadoop architecture. We can achieve multi-node processing which will boost the computation and training will be much faster than the traditional machine. This approach will allow us to work with distributed data processing and since we are trying to implement the logistic regression model from scratch, we can make use of the RDD (Resilient Distributed Dataset) and spark DataFrame.

The spark allows us to create a pipeline and perform all the operations at the time of some action command. While executing the whole pipeline the data is kept into the memory to fast retrieval and once the pipeline is executed the memory is cleared again so that it can be used again or by some other pipeline.

II. Background work

There has been a lot of transformation in the telecommunications industry in recent years carried out by increasing customer expectations, competition, and technological advancement. Mobile devices and the internet is becoming a basic necessity for people in the 21st century. In this time the network providers have two big challenges to face, acquiring new customers and retaining the old ones. In an increasingly dynamic market. One of the main threats to companies is customer churn and which we are addressing in this paper.

A. *The Significance of churn in Telecom*

Customer churn, the phenomenon of subscribers discontinuing services with a telecom provider, is a multifaceted issue with far-reaching implications. High churn rates not only lead to revenue erosion but also necessitate substantial investments in customer acquisition to maintain market share. As the industry witness's constant innovation and the emergence of new players, understanding and mitigating churn have become paramount for the sustainability of telecom businesses.

The telecom sector's unique characteristics contribute to the complexity of churn dynamics. Factors such as intense competition, diverse service offerings, and evolving customer preferences create an intricate landscape were predicting and preventing churn demand sophisticated analytical approaches. Traditional methods often fall short in handling the scale and complexity of telecom datasets, necessitating the exploration of advanced analytics and machine learning techniques.

B. *Objectives of the Study*

The primary objective of this research is to design and implement a predictive model capable of accurately identifying potential churners within a telecom customer base. Through the utilization of PySpark, our approach focuses on handling large-scale datasets efficiently, ensuring scalability to accommodate the dynamic nature of telecommunications data.

C. *Contribution of the Study*

This research contributes to the field by presenting a PySpark-driven approach to telecom churn prediction, emphasizing scalability and efficiency in handling voluminous datasets. The findings of this study aim to empower telecom providers with actionable insights, enabling them to implement targeted retention strategies and thereby reduce churn rates.

III. Theoretical and Conceptual Study

A. *Logistic Regression Overview*

Logistic Regression for Binary Classification: Logistic regression is a statistical method widely employed for binary classification tasks, making it particularly suited for telecom churn prediction where the outcome is binary—whether a customer will churn or not. Unlike linear regression, logistic regression models the probability of the dependent variable belonging to a particular category using the logistic function (sigmoid function).

Logistic Function (Sigmoid): The logistic function, represented as σ(z) = 1 / (1 + e^(-z)), transforms the linear combination of input features (z) into a range between 0 and 1. This transformed output can be interpreted as the probability of the positive class (churn) occurring.

B. *Stochastic Gradient Descent (SGD)*

Optimization Technique: Stochastic Gradient Descent is an iterative optimization algorithm widely used in machine learning for model training. In the context of logistic regression, SGD aims to minimize the logistic loss function by adjusting the weights in the direction that decreases the loss.

Iterative Weight Updates: The key characteristic of SGD is its incremental, iterative approach to updating model parameters. At each iteration, a subset (mini batch) of the training data is used to compute the gradient of the loss function, and the model parameters are updated accordingly. This incremental update makes SGD particularly suitable for large datasets and distributed computing environments.

C. *StandardScaler for Feature Scaling*

Normalization of Feature Vectors: StandardScaler is applied to scale feature vectors before feeding them into the logistic regression model. Feature scaling is essential for algorithms that are sensitive to the scale of input features, ensuring that each feature contributes proportionally to the model's learning process.

Scaling Formula: StandardScaler standardizes the features by subtracting the mean and dividing them by the standard deviation, ensuring that the scaled features have a mean of 0 and a standard deviation of 1.

D. *Implementation in PySpark*

Scalable Distributed Computing: PySpark provides a powerful platform for the scalable and distributed implementation of logistic regression with SGD. By converting the dataset into Resilient Distributed Datasets (RDDs), the logistic regression model can efficiently leverage the parallel processing capabilities of Spark.

Custom Logistic Regression Function: The implementation involves a custom logistic regression function that utilizes SGD for weight updates. The function is designed to handle the distributed nature of the data, making it well-suited for the Spark environment.

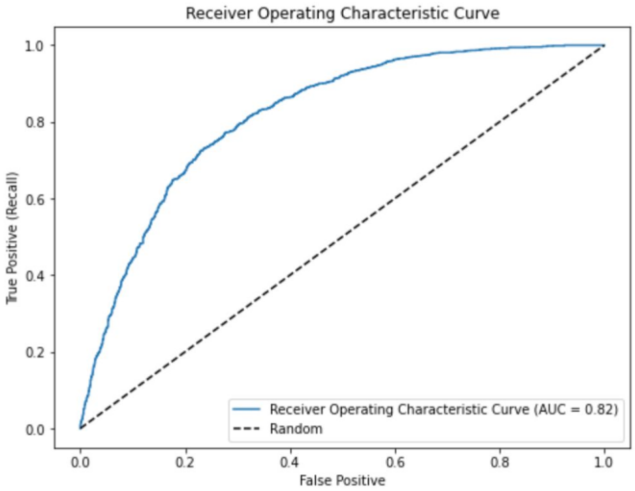
IV. Results and Analysis

A. *Accuracy on Test Data (74.40%)*

The accuracy of 74.40% on the test data indicates the proportion of correctly predicted instances out of the total instances. While accuracy is a commonly used metric, it may not be sufficient for imbalanced datasets or situations where false positives/negatives have different consequences.

B. *Area Under ROC Curve (0.81)*

The ROC curve is a graphical representation of the model's ability to discriminate between positive and negative classes across different thresholds. An AUC of 0.81 suggests a good ability of the model to distinguish between churn and non-churn instances. An AUC closer to 1 indicates better performance.



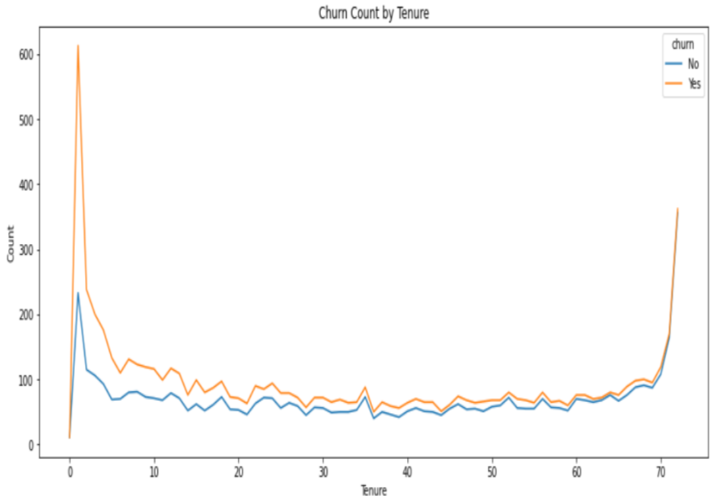
C. *Area Under Precision-Recall Curve (0.78)*

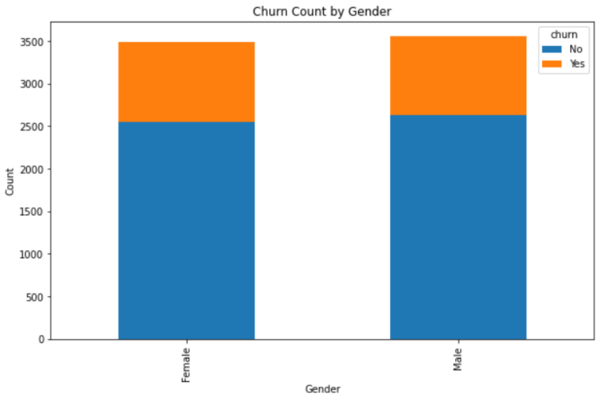
The precision-recall curve focuses on the performance of the model concerning positive instances. An area under the precision-recall curve of 0.78 suggests reasonable precision and recall balance. It indicates that the model can identify true positives while minimizing false positives.

D. *Other Consideration*

1) *Exploratory Data Analysis (EDA):*

Conducting EDA was crucial for understanding the dataset, identifying patterns, and making informed decisions during feature engineering and modeling. It helped us to understand the data better.





2) *Parallel Processing with Spark:*

Leveraging Spark's parallel processing capabilities with RDDs helped us to enhance the efficiency of computation, both for preprocessing as well as for model training.

V. Conclusion

A. *Model Performance:*

The logistic regression model, implemented from scratch using PySpark, demonstrates promising results with a decent accuracy of 74.40%. The AUC values (0.81 for ROC and 0.78 for precision-recall) indicate good discriminative ability, crucial for a churn prediction model.

B. *Insights from Exploratory Data Analysis (EDA):*

1) *Customer Demographics:*

The EDA provided valuable insights into customer demographics, helping us understand the distribution of features such as age, gender, and seniority. These insights could be used for targeted marketing strategies or personalized customer engagement.

2) *Feature Correlation:*

Analyzing feature correlations allowed us to identify relationships between variables. Understanding which features are positively or negatively correlated with churn provides insights into potential drivers of customer attrition.

3) *Data Distribution:*

Examining the distribution of key variables, such as tenure and usage patterns, helped uncover patterns in customer behavior. For instance, long-tenured customers might have different characteristics compared to newer subscribers, influencing their likelihood of churning.

4) *Business Implications:*

Understanding the business context is crucial. The telecom company can use the model to identify potential churners and take targeted actions to retain customers. However, it's essential to weigh the cost of false positives and false negatives based on the business's specific objectives.

5) *Further Evaluation:*

Regular model evaluation and monitoring are necessary to ensure that the model's performance remains robust over time, especially as data patterns may change.

In summary, the implemented model shows promise, but continuous refinement and consideration of business implications are crucial for effective churn prediction and customer retention.

VI. Future Work

While this study has provided valuable insights into telecom churn prediction using PySpark and logistic regression with SGD, there are several avenues for future research and enhancements:

A. *Model Refinement and Feature Engineering*

- Investigate advanced feature engineering techniques to capture more nuanced patterns in customer behavior.
- Explore the incorporation of additional external data sources, such as socio-economic indicators or customer sentiment analysis, to enhance predictive accuracy.
- Experiment with more sophisticated models, including ensemble methods or deep learning architectures, to potentially improve predictive performance.

B. *Scalability and Performance Optimization*

- Assess the scalability of the current approach to handle even larger datasets, potentially incorporating optimizations for distributed computing.
- Investigate parallelization techniques and optimizations within PySpark to further enhance computational efficiency.

C. *Explainability and Interpretability*

- Enhance the interpretability of the model by employing techniques such as SHAP (SHapley Additive exPlanations) values to provide insights into feature contributions.
- Explore the development of a user-friendly interface or visualization tools to assist telecom providers in interpreting and acting upon model predictions.

D. *Real-time Churn Prediction*

- Investigate the feasibility of implementing real-time churn prediction using streaming data, allowing for instantaneous response to changing customer behaviors.
- Assess the integration of the model into live systems, enabling continuous monitoring and adaptation to be evolving patterns.

E. *Comparative Analysis with Alternative Models*

- Conduct a comprehensive comparative analysis with alternative machine learning models, such as support vector machines or random forests, to determine the most suitable algorithm for telecom churn prediction.

F. *Ethical Considerations and Bias Mitigation*

- Examine potential biases in the model predictions and implement strategies to mitigate biases, ensuring fairness and ethical deployment of the churn prediction system.
- Investigate the impact of various preprocessing techniques on bias and fairness in the predictive model.

VII. References

[1]https://github.com/treselle-systems/customer_churn_analysis/blob/master/WA_Fn-UseC_-Telco-Customer-Churn.csv

[2]https://stackoverflow.com/questions/77208295/handling-categorical-missing-data-in-churn-prediction-model-for-telecom-data

[3]https://www.kaggle.com/code/bandiatindra/telecom-churn-prediction