Analyzing Feature Importance in Customer Churn Using Shapley Values

Trần Thiên Anh, Tô Vĩnh An, Nguyễn Minh Chính

26/12/2024

Abstract

This report investigates the factors contributing to customer churn by analyzing feature importance using Shapley values. Logistic Regression was employed to train a predictive model, and SHAP (Shapley Additive Explanations) was utilized to interpret feature contributions. The study uses the Telco Customer Churn dataset, highlighting preprocessing, modeling, and visualization techniques. A production-ready demo was also developed using Streamlit, allowing businesses to explore churn predictions interactively.

1 Introduction

Customer churn, the loss of customers, poses a significant challenge for subscription-based businesses. Understanding churn drivers helps companies design strategies for customer retention. This study employs SHAP values to analyze the importance of features in predicting churn, leveraging Logistic Regression for its interpretability.
Key objectives include:

- Preprocessing and exploring the Telco Customer Churn dataset.
- Training a Logistic Regression model for churn prediction.
- Interpreting feature importance using SHAP.
- Developing a web application for interactive churn analysis.

2 Problem Setup

2.1 Dataset Overview

The **Telco Customer Churn dataset**, sourced from Kaggle [1], includes customer demographics, service details, and churn labels. The target variable, `Churn`, indicates whether a customer discontinued the service. The details of the dataset can be seen in table 1

2.2 Challenges and Solutions

Identifying the factors contributing to customer churn often relies on traditional methods such as analyzing coefficients in Logistic Regression. However, these approaches are often limited by their difficulty in explaining the contribution of each factor or their reliance on linear assumptions. To overcome these challenges, SHAP offers a modern and transparent approach, enabling precise measurement of each factor's impact and helping companies develop more effective marketing strategies.

CustomerID	A unique identifier for each customer	**OnlineBackup**	Indicates if the customer has online backup service
Gender	The gender of the customer.	**DeviceProtection**	Indicates if the customer has device protection service
SeniorCitizen	Indicates if the customer is a senior citizen	**TechSupport**	Indicates if the customer has a tech support service
Partner	Indicates if the customer has a partner	**StreamingTV**	Indicates if the customer has streamingTV service.
Dependents	Indicates if the customer has dependents	**StreamingMovies**	Indicates if the customer has a streamingmovie service
Tenure	The number of months the customer has stayed with the company.	**Contract**	The type of contract the customer has
PhoneService	Indicates if the customer has phone service	**PaperlessBilling**	Indicates if the customer has paperlessbilling
MultipleLines	Indicates if the customer has multiplelines	**PaymentMethod**	The payment method used by the customer
InternetService	The type of internet service the customerhas	**MonthlyCharges**	The amount charged to the customer monthly.
OnlineSecurity	Indicates if the customer has online security service .	**TotalCharges**	The total amountcharged to the customer.

Table 1: Features of dataset

3 Methods

3.1 Preprocessing

Key steps included:

- **Fixing Data Types:** Converted `TotalCharges` to numeric.
- **Handling Missing Values:** Dropped rows with null values.
- **Feature selection:** Dropping feature with low correlation value with churn in figure 1.
- **Data Quality Issues:** Null values and non-numeric columns were cleaned and standardized.
- **Encoding:** Applied one-hot encoding to multi-category features and label encoding for binary features.
- **Standardization:** Scaled continuous features for consistency.
- **Splitting Data:** Divided into 70% training and 30% testing sets.

3.2 Logistic Regression

Logistic Regression models the probability of churn using a logistic function:

$$P(y = 1|X) = \frac{1}{1 + e^{-\beta_0 - \beta_1 X_1 - \cdots - \beta_n X_n}}. \tag{1}$$

Its interpretability and effectiveness in binary classification made it ideal for this task.
The churn prediction model directly impacts Shapley values, with higher accuracy leading to more precise and reliable Shapley values. This helps businesses understand the factors influencing customer decisions and develop effective churn reduction strategies. Logistic regression is a popular model due to its simplicity and ease of implementation in binary classification tasks.

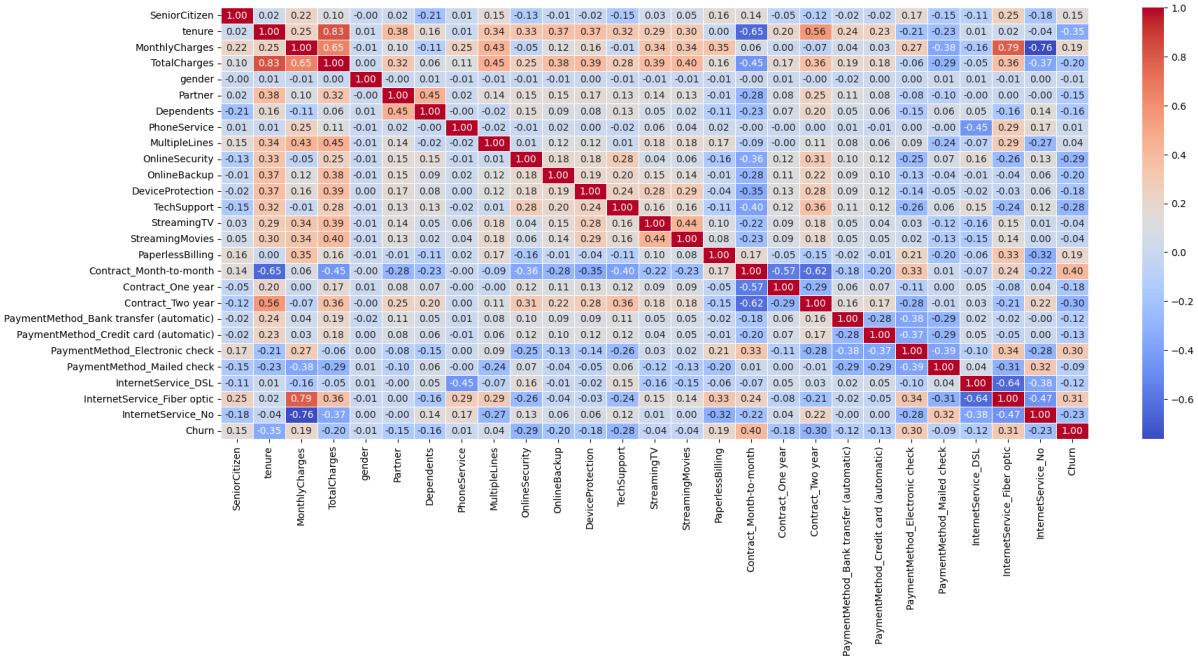

Figure 1: Correlation between features

3.3 SHAP (Shapley Additive Explanations)

SHAP values provide local and global interpretability by quantifying each feature's contribution to the model's predictions. Visualizations such as SHAP summary plots were used to identify key churn drivers.
The SHAP value for a feature i is given by:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)]$$

For comparative purposes, there is a relative scale representing SHAP values. It can be seen at figure 2

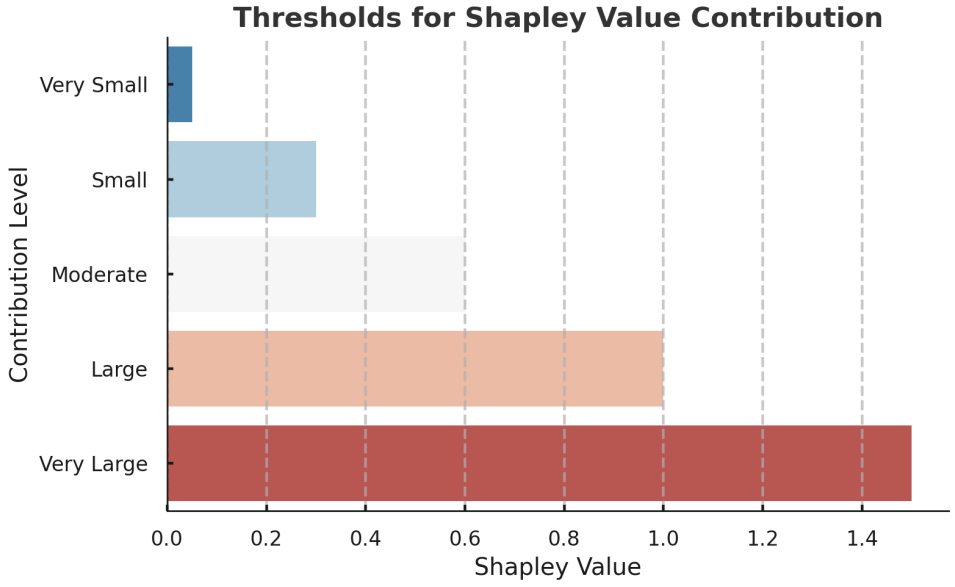

Figure 2: Measurement scale of SHAP value

4 Experiments and Results

4.1 Stratified k-fold Cross-Validation

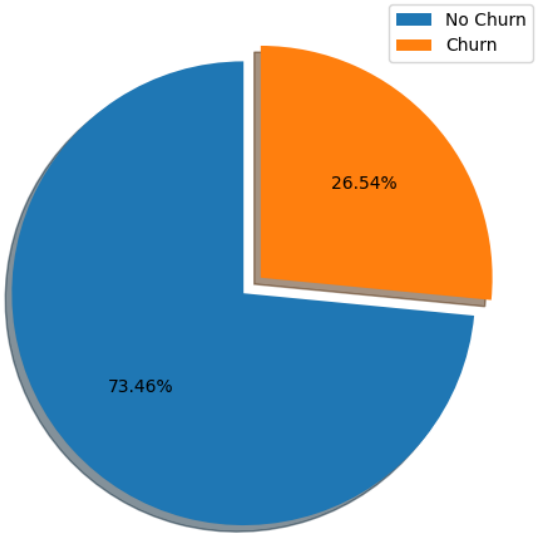

Figure 3: Proportion of customer churn

It can be seen from the pie chart that the ratio between the churn and no churn classes is imbalanced. Therefore, to address this imbalance, the chosen method is stratified k-fold cross-validation. This approach helps in providing a more accurate and reliable evaluation of the model's performance by ensuring that each fold is representative of the overall class distribution.
Stratified k-fold cross-validation is a variation of k-fold cross-validation that ensures each fold is representative of the overall class distribution.
How Stratified k-Fold Cross-Validation Works:

- **Splitting the Data:** The dataset is divided into (k) folds (subsets), ensures that each fold has approximately the same percentage of samples of each target class as the entire dataset.
- **Training and Testing:** The model is trained on (k-1) folds and tested on the remaining fold. This process is repeated (k) times, with each fold being used as the test set exactly once.
- **Averaging the Results:** The performance metrics (e.g., accuracy, precision, recall) are averaged over the (k) iterations

4.2 Model Performance

The Logistic Regression model achieved:

Metric	**Value**
Test Accuracy	0.8128
Cross-Validation Accuracy (Average)	0.8040

Table 2: Performance metrics for Logistic Regression.

The accuracies are quite close (0.8128 vs. 0.8040), indicating that the model's performance is consistent. This suggests that the model is not overfitting to the training data and generalizes well to unseen data.
The slight drop in accuracy from 0.8128 to 0.8040 suggests that the model might perform slightly worse on new, unseen data compared to the training data. However, the difference is small, indicating good generalization ability.

4.3 Feature Importance

SHAP analysis revealed the following top features:

Feature	Shapley Value
Tenure	0.17884
Contract_Month-to-month	0.05472
Contract_Two year	0.0416
...	...
MonthlyCharges	-0.0126
TotalCharges	-0.1180

Table 3: Features sorted based on the SHAP values

Table 3 indicates that the 'Tenure' (0.17884) is the most influential factor in predicting churn. Customers with shorter tenures are more likely to churn. Additionally, contract type plays a significant role, with 'Month-to-month' contracts having a higher churn risk compared to 'Two-year' contracts. The analysis also shows that 'TotalCharges' (-0.1180) and 'MonthlyCharges' (-0.0126) negatively impact churn, with higher charges being associated with lower churn likelihood.

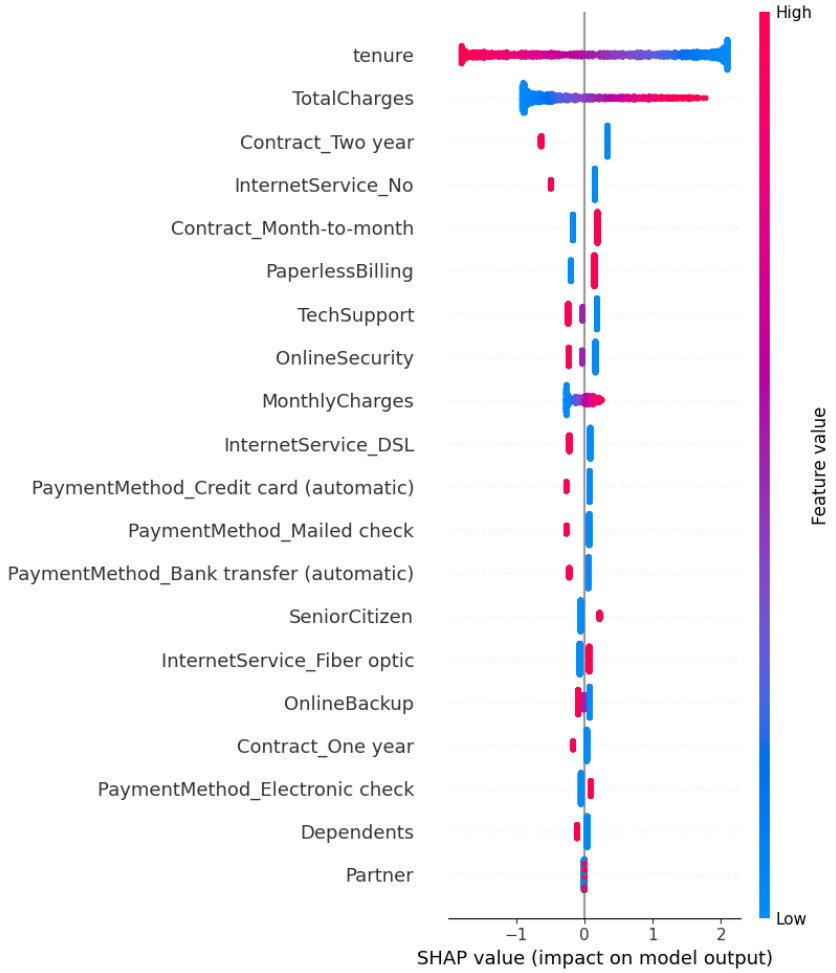

Figure 4: Summary Shapley value

`Tenure` and `TotalCharges` were identified as the most influential features.

5 Production: Streamlit Demo

To demonstrate the results interactively, a web application was developed using **Streamlit**. The application allows users to:

- Upload customer data for churn prediction.
- View SHAP visualizations for feature importance.
- Experiment with hypothetical customer scenarios.

The Streamlit interface provides a user-friendly platform for exploring churn insights and testing retention strategies.

6 Conclusion and Tasks Assigned

6.1 Conclusion

This study highlights the potential of SHAP for analyzing customer churn, with `Tenure` and `TotalCharges` identified as critical factors. The Logistic Regression model achieved reliable accuracy, and the interactive Streamlit demo offers practical utility for businesses.

Implications and Future Work

The successful development and deployment of this system have several implications for businesses. It demonstrates the potential of machine learning and data analysis to drive business decisions and improve customer retention. However, there are areas for future improvement and exploration. For instance, incorporating additional data sources and refining the model could further enhance its accuracy and applicability. Additionally, expanding the system to include other predictive analytics capabilities could provide even more comprehensive support for business strategies.

6.2 Tasks Assigned

- **Tô Vĩnh An:** Responsible for SHAP analysis and visualizations, developing the Streamlit application, and integrating the results.
- **Trần Thiên Anh:** In charge of data collection and writing the technical report.
- **Nguyễn Minh Chính:** Responsible for data preprocessing and building the Logistic Regression model.

Acknowledgments

We thank the contributors of the Telco Customer Churn dataset on Kaggle for their valuable data.

References

[1] Telco Customer Churn Dataset:
`https://www.kaggle.com/datasets/blastchar/telco-customer-churn`

[2] SHAP Documentation: `https://shap.readthedocs.io/`