Introduction to Data Mining Milestone Report

Customer Churn Analysis in Telecom Industry

Yunus Sümer - 1921221010

Erdinç Kuvvet - 2021221057

Yusuf İhsan Şimşek - 2021221023

Contents

Motivation

The problem we are tackling is customer churn in a telecom company. We are specifically focusing on understanding the characteristics and behaviors of customers who churn. The setting we are considering is within the telecommunications industry, where customer retention is crucial, and analyzing factors such as demographic information, service usage, billing details, and customer account information to identify patterns and trends that can help reduce the churn rate and improve overall customer retention.

Method

1. **Data Corrupting:** We deliberately added some missing values and outliers to our initially clean dataset ('data.csv'). The modified dataset is now saved as ('corrupted_data.csv'). This was done to create a more varied dataset for analysis and experimentation.
2. **Data Preprocessing:** We tried handling missing values, converting categorical variables into numerical variables, and scaling numerical variables if necessary. Handling missing values. We handled outliers by removing some of them and scaling rest of them. Also some columns was an object type but has numerical values in it which seemed odd so we converted it to numeric type.
3. **Feature Selection:** The identification of key predictors for customer churn involved a two-step process. Initially, the chi-square method was employed to discern correlated features. Subsequently, the mutual information method was utilized to further refine the identification of relevant features. Based on the results obtained from these methodologies, a judicious selection of features was made, and certain variables were subsequently excluded from consideration.

Preliminary Experiments

Corrupting the Data

1. **Imports and Data Reading:** The necessary libraries, including pandas, numpy, matplotlib, seaborn, and scikit-learn, were imported. The dataset was loaded from the 'data.csv' file.
2. **Faulty Column Correction:** Conversion of the 'TotalCharges' column from a string to numeric, addressing data type discrepancies.Standardization of entries in the 'MultipleLines' and 'OnlineBackup' columns.
3. **Adding Outliers:** Numeric columns ('tenure', 'MonthlyCharges', 'TotalCharges') were converted to float data type. Outliers were introduced to these numeric columns, with a specified percentage of values being randomly replaced with normally distributed outliers.
4. **Adding Missing Values:** A subset of the data containing only 'CustomerID' and 'Churn' columns was preserved. Missing values were introduced across the entire dataset, affecting 20% of the data.
5. **Using Corrupted Data From Now On:** The manipulated dataset was saved as 'corrupted_data.csv' for subsequent analysis.

Data Preprocessing

1. **Data Import**: The dataset was loaded from the 'corrupted_data.csv' file.
2. **Data Inspection:** Initial exploration of the dataset was performed, including a view of the first row, a matrix displaying missing values, and a summary of data information.
3. **Faulty Column Correction:** Correction of data types and replacement of faulty entries. Standardization of entries in columns related to internet and phone services.
4. **Handling Missing Values:** Identification of columns with missing values. Specific handling of missing values in columns like 'tenure,' 'Contract,' 'MonthlyCharges,' and 'TotalCharges.' Imputation of missing values in categorical columns using random sampling based on existing distributions.
5. **Outliers Detection and Handling:** Detection of outliers in 'tenure,' 'TotalCharges,' and 'MonthlyCharges' using the Interquartile Range (IQR) method. Removal of outliers from 'MonthlyCharges.' Winsorization of extreme values in 'TotalCharges.'
6. **Transformations and Encodings:** Conversion of binary categorical columns ('Yes'/'No') to numeric values (1/0). Label encoding for the 'gender' column. Conversion of 'SeniorCitizen' values from float to binary (1/0). Label encoding for 'InternetService,' 'Contract,' and 'PaymentMethod.'
7. **Min-Max Scaling:** Application of Min-Max scaling to 'tenure,' 'MonthlyCharges,' and 'TotalCharges.'
8. **Feature Engineering:** Introduction of a new feature, 'ChurnRiskScore,' based on a calculated metric involving 'MonthlyCharges,' 'Contract,' and 'tenure.'
9. **Identifying Correlated Features:** Chi-square test for categorical features to identify significant associations with 'Churn.' Mutual information method for numerical features to determine significant associations.
10. **Removing Non-Correlated Features:** Removal of features deemed non-significant in the context of predicting 'Churn.'
11. **Output and Error Analysis:** The preprocessed data was saved to a new CSV file ('preprocessed_data.csv'). The impact of each preprocessing step on the dataset was analyzed, including changes in data types, handling of missing values, and the creation of new features. Descriptive statistics and visualizations were used to assess outliers, distribution changes, and the impact of transformations. Error analysis involved checking for inconsistencies, ensuring proper handling of missing values, and confirming the appropriateness of feature encoding.

Next Steps

1. **Model Selection:** We should choose an appropriate machine learning model for our churn prediction task. Considering classifiers such as logistic regression, decision trees, random forests, support vector machines, or gradient boosting etc... Experimenting with multiple models to find the one that performs best for our specific dataset.
2. **Model Training and Evaluation:** We should split the preprocessed data into training and testing sets. Training our selected models on the training set and evaluate their performance on the testing set using metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve. Choosing the model that performs well on these metrics.

Contributions

In this project, each of us contributed equally to the completion of the work. We aim to meet regularly face to face and come together to overcome the problems we face. Whether it was solving complex data or finding the best way to approach a challenge, we brainstormed and worked side by side. Everyone brought their own skills and ideas together, creating a collaborative environment where we could learn from each other. This wasn't just about dividing tasks; it was about being together and making decisions as a team. We are proud of what we achieved because every team member played an important role making this project a true team effort.

References

Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). 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

Shukla, B., Khatri, S. K., Kapur, P. K., Amity University, Amity University. Amity Institute of Information Technology, Computer Society of India., Institute of Electrical and Electronics Engineers. Uttar Pradesh Section, & Institute of Electrical and Electronics Engineers. (n.d.). *2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions) : September 2-4, 2015 : venue, Amity University Uttar Pradesh, Noida, India*.