SYRIATEL CUSTOMER CHURN PREDICTION

DSF-FT4 Phase 3 (Project)

Maureen Kitang’a

1. Business Understanding

1.1 Overview

Customer churn is a significant challenge for businesses, particularly in the telecom industry. Understanding the factors that contribute to customer churn and being able to predict it can help companies develop strategies to retain customers and reduce revenue losses. This project focuses on analyzing the churn dataset from the telecom industry to predict customer churn and gain insights into the key factors driving churn.

1.2 Problem Statement

Syriatel, a telecommunications company, is facing a high churn rate, with many customers discontinuing their services and switching to competitors. The company wants to address this issue by developing a customer churn prediction model. By analyzing the dataset, Syriatel aims to gain insights into factors associated with churn, with the goal of reducing churn rate, increasing customer retention, and improving overall profitability.

1.3 Project Justification

Customer churn can have a significant impact on a telecom company's revenue and market position. By accurately predicting customer churn, the company can take proactive measures to retain customers, improve customer satisfaction, and increase profitability.

1.4 Specific Objectives

The specific objectives of this project are as follows:

- Identify the factors that are most likely to lead to customer churn.
- Develop a model that can accurately predict which customers are at risk of churning.
- Take proactive steps to retain customers who are at risk of churning.

1.5 Research Questions

To achieve the project objectives, the following research questions will be addressed:

- What are the key factors that contribute to customer churn in the telecom industry?
- Can we predict customer churn based on the available customer data?
- Which machine learning algorithms perform well in predicting customer churn?
- How can the telecom company use the churn prediction models to develop effective retention strategies?

1.6 Success Criteria

The success criteria for this project include:

- Developing a robust churn prediction model with high recall score of 0.8
- Identifying the key features and factors that significantly contribute to customer churn.
- Providing actionable insights and recommendations to the telecom company for reducing churn and improving customer retention.
- Demonstrating the value of churn prediction models in enabling proactive retention strategies and reducing revenue losses due to customer churn.

1.7 Project Plan

The project will involve the following essentials:

- A GitHub repository
- Presentation slides for the project
- Cross-Industry Standard Process for Data Mining (CRISP-DM) will be used for this project.

2. Data Understanding

2.1 Overview

In this project, we will work with a customer churn dataset from the telecom industry sourced from Kaggle. The dataset contains information about customers, their usage patterns, and whether they have churned or not.

2.2 Data Description

The dataset contains 3,333 records and spans 21 columns. Of these columns, we identified 4 to be categorical, and 17 as numerical.

Categorical Features:

- `state`: The state where the customer resides.
- `phone number`: The phone number of the customer.
- `international plan`: Whether the customer has an international plan (Yes or No).
- `voice mail plan`: Whether the customer has a voice mail plan (Yes or No).

Numeric Features:

- `area code`: The area code associated with the customer's phone number.
- `account length`: The number of days the customer has been an account holder.
- `number vmail messages`: The number of voice mail messages received by the customer.
- `total day minutes`: The total number of minutes the customer used during the day.
- `total day calls`: The total number of calls made by the customer during the day.
- `total day charge`: The total charges incurred by the customer for daytime usage.
- `total eve minutes`: The total number of minutes the customer used during the evening.
- `total eve calls`: The total number of calls made by the customer during the evening.
- `total eve charge`: The total charges incurred by the customer for evening usage.
- `total night minutes`: The total number of minutes the customer used during the night.
- `total night calls`: The total number of calls made by the customer during the night.

  - `total night charge`: The total charges incurred by the customer for nighttime usage.
  - `total intl minutes`: The total number of international minutes used by the customer.
  - `total intl calls`: The total number of international calls made by the customer.
  - `total intl charge`: The total charges incurred by the customer for international usage.
  - `customer service calls`: The number of customer service calls made by the customer.

3. Data Preparation

3.1 Data Cleaning

The dataset was inspected for missing values, and none were found. This means that our data was complete.

Checking for validity in the data, the dataset was checked for any duplicated values and none were found. For outliers, I chose to check for them in the analysis section.

The phone number column is not important for modeling the data because it contains random numbers and we cannot create dummy values for it. Therefore, we decided to drop the column.

It was observed that the "area code" column, initially of data type int64, only had three distinct types of entries. To better represent the nature of the data and its categorical nature, we made a decision to change the data type of the "area code" column from int64 to object.

3.2 Exploratory Data Analysis

In the Exploratory Data Analysis (EDA) section, we conducted a thorough analysis of the dataset to gain insights into its characteristics and understand the relationships between variables. Here is a summary of what was done:

3.2.1 Univariate Analysis

- We examined the distribution of the "Churn" feature, which indicated that out of the 3,333 customers in the dataset, 483 had terminated their contract, resulting in a churn rate of 14.5%. The data showed an imbalance in the binary classes, which needed to be addressed before modeling.
- The distribution of the "area code" feature was visualized using a pie chart, revealing that almost half of the customers were from area code 415, while the remaining customers were evenly split between area codes 510 and 408.
- The numerical features were analyzed for their distribution using histograms, indicating that most of the features followed a normal distribution. However, the "customer service calls" feature showed multiple peaks, suggesting the presence of multiple modes in the population.
- The distribution of categorical features such as "state," "international plan," and "voice mail plan" was explored using count plots. This analysis revealed insights such as most

customers being from West Virginia, Minnesota, New York, Alabama, and Wisconsin in terms of the "state" feature.

- The "international plan" feature showed that only a small percentage (about 0.1%) of customers had an international plan, while the majority did not.
- Similarly, for the "voice mail plan" feature, about 0.3% of customers had a voicemail plan, while the majority did not.

3.2.2 Bivariate Analysis

- The relationship between variables was explored using box plots and count plots. For example, the box plot showed that customers who terminated their accounts were primarily from area codes 415 and 510, and there were some outliers in the "customer service calls" feature.
- The distribution of categorical features based on the churn rate was examined. It was observed that customers from Texas, New Jersey, Maryland, Miami, and New York had a higher churn rate. Additionally, customers without an international plan or voicemail plan had a higher likelihood of churning.

3.2.3 Dealing with Outliers

- Outliers can impact predictive models by introducing noise or skewing the training process. Therefore, we applied a method to remove numerical outliers beyond 3 standard deviations from the mean. After removing the outliers, the dataset contained 3,169 records.

3.2.4 Features Correlation

- We computed the correlation matrix and plotted a heatmap to identify features that exhibited high correlation with the target variable, "churn." It was found that features such as "total day charge" and "total day minutes" were fully positively correlated, as were "total eve charge" and "total eve minutes," "total night charge" and "total night minutes," and "total int charge" and "total int minutes." This correlation is expected since the charge is directly dependent on the corresponding minutes used.

3.2.5 Multicollinearity check

- Multicollinearity occurs when features in the dataset are highly correlated with each other, leading to issues during modeling. To address this, we calculated the correlation matrix and identified highly correlated features with a correlation value above 0.9. These features were dropped from the dataset to mitigate multicollinearity effects.

3.3 Feature Engineering

Several feature engineering techniques were applied to the dataset. Here is a summary of what was done and the reasons behind each step:

1. Label Encoding: The categorical variable 'churn' with values 'yes' and 'no' is converted into numerical values using label encoding. Label encoding assigns a unique integer to each category ('yes' becomes 1 and 'no' becomes 0). This conversion allows machine learning algorithms to work with categorical data, as they generally require numerical inputs.
2. One-Hot Encoding: Several categorical variables including 'state', 'area code', 'international plan', and 'voice mail plan' are converted into a set of binary features using one-hot encoding. This technique creates a new feature for each category and assigns a value of 1 if the category is present, and 0 if it is not. One-hot encoding is used to prevent the model from assuming any ordinal relationship between the categories, as it could lead to incorrect interpretations.
3. Scaling the Data: Numeric features are scaled using Min-Max normalization. Scaling is performed to bring all the features to a comparable range and prevent certain features from dominating the model due to their larger magnitude.

4. Modeling

To begin, we point out that the dataset is imbalanced, with only 14.5% of the data being classified as "churn" and the rest being classified as "not churn."

Our baseline model was LogisticRegression , just to test how our chosen attributes would perform on a basic level. We then explored other more sophisticated models like the RandomForest, Decision trees, and XGBoost.

4.1 LogisticRegression

The logistic regression model returned a recall score of 74%. Though this was not our desired recall score, it was good for a baseline model and gave us a good starting point for our other models.

The confusion matrix evaluation showed that the model had a higher number of true positives and true negatives than false positives and false negatives. This indicates that the model is making correct predictions more often than incorrect ones and is not overfitting.

According to the model, total day charge, customer service calls, total eve charge were the top three most important features.

4.2 Decision Trees

The decision tree model returned a recall score of 73%, which was good but not better than our baseline model.

The confusion matrix evaluation showed that the model had a higher number of true positives and true negatives than false positives and false negatives. This indicates that the model is making correct predictions more often than incorrect ones and is not overfitting.

According to the model, total day charge, total eve charge, total intl charge were the top three most important features.

4.3 Random Forest

The random forest model returned a recall score of 74%, which was the same score as our baseline model but not our desired score.

The confusion matrix evaluation showed that the model had a higher number of true positives and true negatives than false positives and false negatives. This indicates that the model is making correct predictions more often than incorrect ones and is not overfitting.

According to the model, total day charge, total intl calls, total eve charge were the top three most important features.

4.4 XGBoost

The XGBoost model returned a recall score of 77%, which was better than all the previous models but not our desired score.

The confusion matrix evaluation showed that the model had a higher number of true positives and true negatives than false positives and false negatives. This indicates that the model is making correct predictions more often than incorrect ones and is not overfitting.

According to the model, total day charge, total intl calls, number vmail messages were the top three most important features.

5.Model Evaluation

5.1 Models' Comparison

XGBClassifier has the highest recall score, followed by RandomForestClassifier and, LogisticRegression. The DecisionTreeClassifier has the lowest recall score of 0.73.

The ROC curve analysis shows that the XGBClassifier has the best performance, followed by the RandomForestClassifier, DecisionTreeClassifier, and LogisticRegression. The XGBClassifier has the highest AUC score of 0.884, while the LogisticRegression has the lowest AUC score of 0.79.

5.2 Model Tuning

Based on the evaluation of the models using recall scores and ROC AUC, it is observed that the XGBoost classifier and the RandomForest classifier have shown promising performance. To further improve their performance, model tuning can be performed using GridSearch.

5.3.1Tuning RandomForest

To accomplish this, we utilized GridSearchCV, a powerful tool for hyperparameter tuning. We defined a hyperparameter grid containing various options for parameters such as *max_depth*, *n_estimators*, *min_samples_split*, and *criterion*. The grid represented different combinations of these parameters that were tested to find the optimal configuration.

After performing the grid search with cross-validation, we obtained the best parameters for our Random Forest classifier.

Next, we instantiated a new Random Forest classifier using the tuned hyperparameters and fitted it to the resampled training data. The tuned model was then used to make predictions on the test data.

The evaluation of the tuned Random Forest model demonstrated significant improvements compared to the baseline logistic regression model. The model returned a recall score of 0.76.

The confusion matrix further supported the model's effectiveness, as it showed a higher number of true positives and true negatives compared to false positives and false negatives. This indicates that the model is not biased towards any class and avoids overfitting.

Additionally, we analyzed the receiver operating characteristic (ROC) curve and computed the area under the curve (AUC) score. The ROC curve demonstrated good discriminatory power of the model, with an AUC score of 0.86. This score indicates that the model can effectively distinguish between positive and negative instances.

5.3.2 Tuning XGBoost

The hyperparameter grid was defined with various options for parameters such as *learning_rate, max_depth, min_child_weight, subsample, and n_estimators*. These parameters represent different aspects of the XGBoost model that were tested to find the best configuration.

After performing the grid search with cross-validation, we obtained the best parameters for our XGBoost classifier.

Next, we instantiated a new XGBoost classifier using the tuned hyperparameters and fitted it to the resampled training data. The tuned model was then used to make predictions on the test data.

The evaluation of the tuned XGBoost model showed significant improvements compared to the baseline logistic regression model. The classification report indicated an accuracy of 93%, precision of 0.80, recall of 0.79, and an F1-score of 0.79. These metrics indicate that the model performs well in correctly classifying both positive and negative instances.

The confusion matrix further supported the effectiveness of the tuned XGBoost model, as it showed a higher number of true positives and true negatives compared to false positives and false negatives. This suggests that the model avoids overfitting and does not exhibit a bias towards any specific class.

Furthermore, we analyzed the receiver operating characteristic (ROC) curve and computed the area under the curve (AUC) score. The ROC curve demonstrated good discriminatory power of the model, with an AUC score of 0.88. This indicates that the model can effectively distinguish between positive and negative instances.

In conclusion, the tuned XGBoost model exhibits improved performance compared to the baseline logistic regression model. It achieves higher accuracy, precision, recall, and demonstrates good discriminatory power. These results indicate that the tuned XGBoost classifier is a suitable model for the binary classification task and can provide reliable predictions for the given dataset.

6. Conclusion

The recall score of our XGB classifier was 79%. While this is still a good predictive model, we would like to undertake further feature engineering to boost this recall score if we had more time. We achieved our objectives to be able to predict customer churn and had an acceptable recall score.

Recommendations

- Offer discounts or promotional offers to customers in area code 415 and 510, as these areas have a higher churn rate. This can help incentivize customers to stay with the company.
- Improve customer service quality and reduce the number of customer service calls. Enhance training programs for customer service representatives to ensure prompt and effective resolution of customer issues, leading to higher customer satisfaction and reduced churn.
- Evaluate the pricing structure for day, evening, night, and international charges. Consider adjusting pricing plans or introducing discounted packages to address the higher charges associated with customers who churn.
- Focus on customer retention strategies in states with higher churn rates, such as Texas, New Jersey, Maryland, Miami, and New York. This can involve targeted marketing campaigns, personalized offers, or improved customer support tailored to the specific needs and preferences of customers in those states.

- Enhance the value proposition of the voicemail plan to increase adoption among customers. Highlight the benefits and convenience of voicemail services, and consider offering additional features or discounts to encourage customers to sign up.