Customer Churn Prediction for Syriatel,Report.

Author: Caroline Wambui Kimani

Date: 8th December 2024

BUSINESS UNDERSTANDING

SyriaTel is facing a challenge of customer retention in a competitive telecommunications market, where churn leads to revenue losses and this has an impact in its market reputation. The goal is to develop a predictive model to identify customers at risk of churning, that will enable the company to proactively implement targeted retention strategies and reduce churn. The insights derived from this analysis will be crucial for the stakeholders such as Customer Retention Team, Marketing and Sales Team, Customer Support Team, Business Executives and Decision Makers in further understanding the factors influencing customer churn.

DATA UNDERSTANDING

The dataset was sourced from Kaggle.

The dataset contains information on 3,333 customers of the telecommunications company "SyriaTel." It features 20 attributes, including customer demographics, call usage patterns during different times of the day, and subscription details such as voice mail and international plans. The account length attribute indicates how long a customer has been with SyriaTel, offering a valuable measure of their lifetime value.

The dataset features can further be outlined as shown below:

- state: the state the customer lives in
- account length: the number of days the customer has had an account
- area code: the area code of the customer
- phone number: the phone number of the customer

- international plan: true if the customer has the international plan, otherwise false
- voice mail plan: true if the customer has the voice mail plan, otherwise false
- number voicemail messages: the number of voicemails the customer has sent
- total day minutes: total number of minutes the customer has been in calls during the day
- total day calls: total number of calls the user has done during the day
- total day charge: total amount of money the customer was charged by the Telecom company for calls during the day
- total eve minutes: total number of minutes the customer has been in calls during the evening
- total eve calls: total number of calls the customer has done during the evening
- total eve charge: total amount of money the customer was charged by the Telecom company for calls during the evening
- total night minutes: total number of minutes the customer has been in calls during the night
- total night calls: total number of calls the customer has done during the night
- total night charge: total amount of money the customer was charged by the Telecom company for calls during the night
- total intl minutes: total number of minutes the user has been in international calls
- total intl calls: total number of international calls the customer has done
- total intl charge: total amount of money the customer was charged by the Telecom company for international calls

- customer service calls: number of calls the customer has made to customer service
- churn: true if the customer terminated their contract, otherwise false

DATA PREPARATION

The data was loaded after importing the relevant libraries, the data was then checked for duplicates, missing values and any irrelevant columns that may hinder the accuracy of the data prediction. Data information and description was done as well to further understand the data.

DATA CLEANING

Cleaning the data ensures reliability and accuracy of the data, therefore the data had no missing values and duplicates. It was clean data, however the 'churn' column had to be converted to a numerical value of 0 and 1 so as to be compatible with the machine learning algorithms. No renaming of columns was done in this data however the data types found in the data check had to be confirmed.

EXPLORATORY DATA ANALYSIS [EDA].

Visualizations and statistical summaries were used to explore relationships:

- **Churn Distribution**: It was observed that around 14.5% of the customers had churned, that is 483 customers out of the 3333 customers registered.
- **Area Code Distribution**: Area code 415 had the highest number of customers leading by 1655 followed by area 510 which was also closely followed by area 408 with a registered customer base of 840 and 838 respectively. This distribution was also checked for outliers and it was discovered that customers who are likely to churn may be from area codes 510 and 415.

- **Distribution of Features:** Churn column was not included in the analysis of this distribution. The features were checked for outliers and features such as minutes or calls showed high usage by customers. It was mostly a right skewed distribution meaning only few customers make a lot of calls or use a lot of minutes.
- **Distribution by International Plans:** The findings were that only 323 customers had an international plan while 3010 customers did not have an international plan making up a 10% and 90% distribution rate respectively.
- **Voice Mail Plan Churn Rate:** Only 922 customers had a voicemail plan and 2411 customers did not have a voicemail plan making up a 28% and 72% rate respectively.
- **Total Minutes by Category:** The findings were that total evening minutes, total night minutes and total day minutes made up the largest part of the total minutes used by customers and total international calls were the least favoured making up a 1.7%. By identifying the preference of the customers, it's easy to design tariffs that satisfy the experience and the needs of the customer.
- **Customer Service and Customer Churn Rate**: Churn rate increased with more customer service calls; this could suggest that the frequent callers were not always satisfied with the level of issue resolution hence the churn.
- **Correlation Matrix:** There were strong positive correlations between minutes and charges and international usage and charges. Moderate correlation between voice mail messages and customer care services calls. There was no correlation between account length and phone numbers. Therefore, from the dataset it was established that the correlation of relationships was negative.

  Using tools like Seaborn and Matplotlib, visualizations such as bar plots and histograms revealed these patterns.

MODELLING

To improve model performance, features were refined:

- Focused on independent variables such as Total day minutes, Total evening minutes, Total night minutes and Total international minutes to determine the multicollinearity.
- A variance calculation factor used to assess each feature.
- Columns with high correlation were dropped and the data types for the remaining data was checked.
- One hot encoding was applied to the dataset for further preparation of modelling and phone number column was dropped and data set information was lastly verified for further modelling to expedite accuracy.

The data was split into X and Y for training and testing, followed by scaling and the data was checked for any imbalance that may hinder accurate predictions. SMOTE was used to train the data and help balance the data.

Two machine learning models were trained to predict churn:

- **Logistic Regression**: A simple yet effective baseline model.
- **Decision Tree**: A more complex model capable of capturing non-linear relationships.

EVALUATION METRICS:

- **Accuracy**: Measures overall correctness but can be misleading with imbalanced datasets.
- **Precision and Recall**: Focused on churn prediction effectiveness.
- **AUC (Area Under Curve)**: Showed the model's ability to distinguish between churners and non-churners.
-

FINDINGS:

- Logistic Regression provided a good balance between precision and recall.
- Decision Tree slightly outperformed Logistic Regression in AUC but was less consistent across precision and recall.

RECOMMENDATIONS

- **Retention Programs**: Focus on at-risk customers (e.g., short-tenure customers) by offering personalized discounts or loyalty rewards.
- **Flexible Contracts**: Encourage longer-term commitments through incentives like discounts for annual plans.
- **Enhanced Customer Support**: Address complaints about pricing or service reliability promptly to reduce churn.

Key Drivers of Churn:

- Customers with high monthly charges or short tenures were more likely to churn.

CONCLUSION

This project was an excellent introduction to real-world data science applications. I learned how to handle real-world datasets, derive insights, and build predictive models. While the models showed decent predictive power, the imbalanced dataset influenced performance, and future work could involve experimenting with advanced techniques like XGBoost or oversampling.

NEXT STEPS

1. **Address Imbalance**: Resample the data or use techniques like SMOTE to balance the dataset.
2. **Advanced Models**: Test algorithms like Random Forest or Gradient Boosting for improved predictions.
3. **Business Application**: Build a dashboard to visualize real-time churn predictions for actionable insights.

4. **Continuous Updates**: Keep refining the model with new data to ensure relevance over time.