

1 Branch

0 Tags

Go to file

t

Go to file

+

Add file

Code

lauramwichekha

ReadMe updated

14371fd · 11 minutes ago

<div></div> .ipynb_checkpoints	add changes	22 minutes ago
<div></div> Notebook.pdf	add changes	22 minutes ago
<div></div> Project Presentation.pptx	ReadMe updated	11 minutes ago
<div></div> README.md	Final Project	1 hour ago
<div></div> Student.ipynb	add changes	22 minutes ago
<div></div> Syria tel data.csv	Final Project	1 hour ago

README

# SyriaTel Customer Churn Prediction

## Project Task

Build a binary classification model to predict whether a SyriaTel customer will churn (i.e stop doing business) in the near future.

## Business Understanding

The telecom provider SyriaTel wants to minimize revenue loss from client attrition. SyriaTel can enhance customer satisfaction and retain consumers by implementing focused retention methods aimed at identifying high-risk clients.

## Business Objectives

1. Determine which customers are at danger of churning: Make an accurate guess as to which ones are most likely to cut off service.

2. Put focused retention methods into practice: Create specialized marketing strategies to speak to the unique requirements of at-risk clients and persuade them to stick with SyriaTel.
3. Boost client contentment Improve the whole customer experience to lower attrition and increase steadfast loyalty.
4. Efficient resource allocation can be achieved by concentrating retention efforts on customers who are most likely to leave.

## Data Mining Objectives

---

Create a categorization model using pertinent customer qualities and past data that can reliably forecast client attrition. It should be possible for the model to distinguish between clients who are likely to churn and those who are not.

## Data Loading And Understanding

---

This project makes use of the SyriaTel dataset, which can be found in SyriaTel Data.csv. The following columns and their meanings are as follows:

1. State: this usually indicates the customer's geographic location
2. Account Length: this indicates the duration that the customer has held their account
3. Area Code: this is the customer's phone number
4. Phone Number: this is typically a unique identifier for customers
5. International Plan: this is a binary variable that indicates whether the customer has an international calling plan
6. Voice Mail Plan: This is a binary variable that indicates whether the customer has a voicemail plan
7. Number Vmail Messages: this is the number of voicemail messages that the customer has received
8. Total Day Minutes, Total Day Calls, Total Day Charge: These columns reflect the customer's usage during the day
9. Total Eve Minutes, Total Eve Calls, Total Eve Charge: These columns reflect the customer's usage in the evening
10. Total Night Minutes, Total Night Calls, Total Night Charge: These columns reflect the customer's usage at night
11. Total Intl Minutes, Total Intl Calls, Total Intl Charge: These columns correspond to the customer's usage abroad
12. Customer Service Calls: The quantity of calls made by the customer
13. Churn: This is the target variable that indicates whether a customer has churned (1) or not (0); all other columns are potential features for modeling.

During our EDA, there are some findings that have come up and need to be addressed before getting into modeling. Below are the findings:


Finding 1: Converting data type


Upon inspecting the data, we identified an important data type issue related to the 'area code' column. Although it is represented as an integer in the dataset, the values it contains are essentially placeholders or labels, not numerical values that carry mathematical significance. To prevent potential interference with our predictive modeling process, we have undertaken the step of converting this column to a string data type. By doing so, we ensure that the 'area code' column is treated as a categorical variable with no numerical significance. This transformation aids in maintaining the integrity of our predictive model, especially when the model relies on numerical inputs, preventing any misinterpretation of the 'area code' as a quantitative feature.

#### Finding 2: High correlation - Multicollinearity

Our examination of the heatmap representation of the data revealed that several columns exhibit high levels of correlation with each other. This observation indicates the presence of multicollinearity, a condition where independent variables in our dataset are highly interrelated. Multicollinearity can make it challenging to discern the unique impact of each independent variable on the dependent variable in our modeling. This issue has the potential to create overfitting problems, particularly when employing Logistic Regression, as this method is sensitive to multicollinearity.


To address this concern, it will be crucial to take specific steps to handle multicollinearity in our modeling process. These steps might involve employing techniques such as feature selection, dimensionality reduction, or regularization to mitigate the adverse effects of multicollinearity. By doing so, we can enhance the reliability and interpretability of our models and ensure that they provide accurate insights into the relationships between our independent and dependent variables.

Heatmap to check how the columns are correlated heatmap  [alt text](#)

Histograms for numerical features to visualize their distribution histogram  [alt text](#)

#### Finding 3: Outliers

In our analysis of the data, we observed the presence of a significant number of outliers in our dataset, as indicated by the boxplots. These outliers have the potential to impact our modeling process. However, it is important to note that, in this case, these outliers are not anomalies that should be removed. Instead, they are a noteworthy aspect of our dataset that we should be aware of during our modeling process. These outliers may carry valuable information or insights that could be relevant to our analysis, and, as such, it is essential to consider and account for them when developing our models and interpreting the results. Understanding the nature and impact of these outliers is a critical part of ensuring the robustness and accuracy of our data analysis.

Box plots to identify outliers and visualize the spread of data  [alt text](#) boxplot

#### Finding 4: Scatter plots showing multicollinearity

Our analysis of the scatter plots has revealed the presence of features that exhibit a perfect correlation with each other. This perfect correlation is a clear indicator of multicollinearity, a condition where independent variables in our dataset are highly interrelated, possibly to the extent that they move in perfect synchronization. Multicollinearity poses a significant challenge to our modeling process and can lead to less reliable statistical inferences.

To address this concern, it is imperative that we take appropriate measures to mitigate the impact of multicollinearity in our model. Strategies for handling multicollinearity may include feature selection, dimensionality reduction, or regularization techniques. By applying these methods, we can improve the robustness and interpretability of our model, thereby ensuring that it delivers trustworthy and accurate statistical insights while addressing the challenges posed by multicollinearity.

scatterplot

### 3. Data Preparation

#### step 1: Removing irrelevant columns

Given that our dataset does not contain any missing values and duplicates, the next step is to streamline our data by removing columns that are not essential for our analysis or modeling.

#### Step 2. Feature Engineering

In our dataset, we have identified that both our target variable and certain feature columns are categorical in nature. To effectively use this data in our modeling process, it is advisable to encode these categorical variables into a numerical format.

#### Finding 5: Encoding

Encoding categorical variables is a crucial step, as many machine learning algorithms require numerical inputs for model training.

By performing appropriate encoding techniques, such as one-hot encoding or label encoding, we can convert the categorical values into a numerical representation that the model can understand. This encoding process will facilitate the modeling phase and ensure that our machine learning algorithms can effectively utilize the information contained within these categorical columns, leading to more accurate and reliable model predictions.

#### Finding 6: Imbalanced data

The 'churn' column represents a binary outcome, where 'False' represents customers who did not churn (i.e., stayed), and 'True' represents customers who did churn (i.e., left). The 'churn\_encoded' column is a numerical representation of 'churn,' where 0 typically corresponds to 'False,' and 1 corresponds to 'True.' The majority of the data (about 85.51%) falls into the 'False' or 0 category, while the minority of the data (about 14.49%) falls into the 'True' or 1 category. This indicates that the dataset might be imbalanced, with a higher proportion of non-churned customers.

#### Step 3. Choosing the Target and the Features

We've chosen the relevant features and the column churn as our target for our models.

### 4. Modeling

Develop a predictive model designed to anticipate whether a customer is on the verge of discontinuing their engagement with Syria. The primary objective is to curtail financial losses stemming from customers who have a short-lived association with the entity.

#### Model 1: Logistic Regression

## Step 1. Address the findings above

Before proceeding with the development of our Logistic Regression model, it's important to address the previous findings, as they can significantly impact the model's accuracy and reliability:

### 1. Handling Multicollinearity

Some columns like 'total day minutes' and 'total day charge' that are perfectly correlated columns are providing identical information, and keeping both in the dataset doesn't provide any additional value to your analysis or modeling. It's more straight forward to include the 'total day minutes' variable, as it represents the customer's actual behavior (minutes used) rather than a derived value (charge). This applies to 'total day charge', 'total eve charge', 'total night charge', 'total intl charge' as well.

### 2. Handling Class Imbalance with SMOTE

The majority of the data (about 85.51%) falls into the 'False' or 0 category, while the minority of the data (about 14.49%) falls into the 'True' or 1 category. This indicates that the dataset might be imbalanced, with a higher proportion of non-churned customers. In this case we will use SMOTE (Synthetic Minority Over-sampling Technique) to address the issue. This Increase the number of instances in the minority class by creating duplicates or generating synthetic examples.

## Step 2. Scale the Data

This is to ensure that the features are on a common scale.

## Step 3. Perform a train-test split

Split the dataset into training and testing sets. This separation ensures that you have a clear distinction between the data used for training the model and the data used for evaluating its performance.

## Step 4. Build and evaluate a baseline model, Decision Trees Model and KNN Classifier Model

After applying SMOTE, we proceed to train a machine learning model on the resampled training data using `X_resampled` and `y_resampled`. In this case, we will use a logistic regression model. KNN is a non-parametric and instance-based learning algorithm, meaning it doesn't make underlying assumptions about the data distribution and makes predictions based on the similarity between instances.

### 5. Evaluation

The logistic regression model shows a balanced performance with reasonably good accuracy, precision, recall, and F1 score. It captures positive cases effectively while maintaining precision. The ROC AUC score is also decent.

The decision tree model exhibits high accuracy, especially for the majority class. However, it has lower precision, recall, and F1 score for the minority class. This suggests that it may not perform as well on classifying the minority class. The model is well-suited for imbalanced datasets.

The KNN classifier model achieves decent accuracy and performance for the majority class but struggles with the minority class, similar to the decision tree model.

### 6. Model of Choice: Decision Trees

From the above models, We've chosen to use the Decesion trees Model.

The decision tree model was evaluated with precision, recall, and F1 score for both Class 0 and Class 1. From the accuracy results above our model correctly predicts the class labels for the majority of instances in the test data. The precision metric is very important as it measures how accurate the model is at identifying the majority class which is the customers who don't churn. The downside of our model is that it has lower precision, recall, and F1 score for the minority class. This suggests that it may not perform as well on classifying the minority class, the same way it does the majority class. But this is majorly attributed to the fact that our data is also highly imbalanced. But compared to the other two models, this one performs much better. Our model performs well in terms of precision and recall for both classes on the holdout test data, thus it can be deployed in a real-world scenario.

## 7. Recommendations and Future Investigations

**Customer Service Calls Investigation:** Dig deeper to understand why some customers need to contact customer service frequently. This will help in finding ways to better assist them.

**International Plan Churn Investigation:** Since some of the customers with international plans are leaving, it's essential to explore ways to retain these customers.

**High Churn States Analysis:** Look into the states where many customers are leaving to identify any patterns or reasons for the high churn rates.

**Incentives for High Bill Customers:** Find ways to encourage customers with high daily charges (over \$55) to stay with SyriaTel. This might involve offering extra benefits and perks. Currently, all of these high-bill customers are leaving, which is a concern.

**Incentives for Customers who stay more than 6 months:** Find ways to encourage customers to stay with the company even longer, eg giving them loyalty points, offers, etc, as this will help in creating a form of loyalty.

## Conclusion

In conclusion, the decision tree model appears to be providing reasonably good results, especially for the



## Releases

No releases published

[Create a new release](#)

## Packages

No packages published

[Publish your first package](#)

## Languages

● Jupyter Notebook 100.0%