June 14, 2024

Talal Majrad

ID: 011396024

C964: Computer Science Capstone

Letter of Transmittal

6/14/2024

James Williams

TeleNet

123 Plymouth Rd, MI 48201

Dear James Williams,

I'm writing this letter to present a solution to a very critical issue that our company and many other companies face. This issue is known as customer churn. Churn is the rate at which a business loses customers or subscribers. The loss of customers significantly impacts the revenue and growth of the company. It is in the best interest of the company to maintain a relationship and business with as many clients or customers as possible. So why not do something about it, why lose out on revenue when action can be taken to counter the impact?

Here is my solution to this problem. I present to you a Customer Churn Prediction Model. The ability to predict which customers are likely to end their relationship with our business, and take proactive measures to retain them. This has substantial financial benefits for the company, potentially millions can be saved by keeping customers engaged.

The proposed solution is centered around a machine-learning application built to predict customer churn. This application will analyze customer data with various data points, such as age, tenure, usage frequency, support calls, payment delay, subscription type, and contract length. The model identifies patterns and trends and predicts the likelihood of each customer churning. With the help of these predictions, our customer retention team can add those customers to the at-risk customers with personalized interventions.

The financial benefits of this application include increased customer retention, cost savings, improved customer experience, and data-driven decisions. Retaining existing customers is less expensive than acquiring new ones, and this model will help optimize our marketing and customer service efforts. The insight provided by the model will enable the company to make informed decisions about customer retention strategies.

The implementation plan is straightforward. The primary costs will involve the initial setup of the model, both hardware and software, labor costs, and environmental costs. The estimated costs are as follows:

- $10,000 - Initial setup of the model (including server infrastructure, software licenses, and development tools)
- $120,000 - Labor costs (Data Scientists, Developers, and Project Manager for four months)
- $24,000 - Environmental costs (deployment, hosting, and maintenance for the first year)

These costs are insignificant compared to the long-term savings from the application. The timeline for the project consists of five phases.

1. 2 months – Data collection and model development
2. 1 month – Model testing and validation
3. 1 month – Implementation and staff training
4. 2 weeks – Team training
5. Ongoing – Monitoring the model

As the lead data analyst, I have extensive experience in machine learning and data analysis. My background includes developing predictive models and implementing data-driven solutions that have significantly improved customer satisfaction in previous roles.

In conclusion, the customer churn prediction model is a strategic investment for our company's future. By proactively addressing customer churn, we can enhance retention, reduce costs, and ultimately promote growth. I am confident that this project will deliver substantial benefits and look forward to your approval to proceed.

Sincerely,

Talal Majrad - Customer Churn Prediction Application

Project Proposal Plan

Project Summary

Problem Description

Customer Churn is when clients or subscribers discontinue their relationship with a company. This problem presents major revenue loss to our company, and the ability to predict which customers are most likely to churn lifts the risk of company losses. With the predictions provided, proactive measures to retain customers could provide substantial financial benefits to the company, potentially saving millions by keeping customers engaged and satisfied.

Clients and their Needs

Understanding the needs of clients and customers is the way to a flourishing company. Clients and customers want to be satisfied with the service provided. The prediction application I present will help to understand our clients and customers better. Analyzing the data allows us to predict which customers are most likely to end their relations with our company. This gives us the chance to take proactive measures targeted towards high-risk customers. The measures to be taken can keep the customer satisfied and continue our relations.

Deliverables

- Churn Prediction Model: A machine learning model designed to predict customer churn based on various data points such as tenure, usage frequency, support calls, payment delay, subscription type, and contract length.

- User Interface: An interactive dashboard that allows users to pass customer datasets and receive churn predictions.

User Guide for Customer Churn Prediction Application

Welcome to the Customer Churn Prediction Application! This guide will help you set up and run the application, and understand the results it provides. You can run the Customer Churn Prediction Application using platforms like Google Colab or Jupyter Notebook. Both platforms provide an interactive environment for running machine learning applications.

First, ensure you have access to either Google Colab or Jupyter Notebook and upload the provided application code file to the platform you are using.

For those planning to use Jupyter Notebook, you can install it by running the command “pip install notebook” in the command terminal. To install the required libraries run this command in the terminal `pip install pandas numpy matplotlib seaborn plotly scikit-learn ipywidgets termcolor`. Once installed, you can then run the command “Jupyter Notebook”. This will open a Notebook.

For those using Google Colab, you can search Google Colab on Google and you can directly open a notebook. For Google Colab these libraries are pre-installed so you can get started right away.

The attached CSV file, which will be submitted with the project, can be directly used once the file is uploaded. If you wish to use a different dataset, you will need to modify the code. Specifically, locate the line where the dataset file is being read and change the file name in the code to match the name of your uploaded dataset file. For example, replace `df = pd.read_csv('customer_churn.csv')` with the actual name of your dataset file. The new dataset must follow the same format as the attached file, meaning it should have the same columns, as the code was built with the original example dataset in mind.

Next, run all the cells at once using the toolbar or separately run each cell in sequential order. This will perform the required functions such as the preprocessing of data as well as the actual training of the model and the prediction. The application will create different types of visuals based on the information to assist you in understanding the data. These include the data distribution visuals – bar charts to show the distribution of various features like customers based on tenure, total spending, payment delay, etc. Towards the end of the code, there will be a section that applies the Decision Tree algorithm to a scatter plot. This scatter plot gives us insight into how the Decision Tree algorithm works, by displaying the predictions and dividing the possibilities of churn or no churn between red and green colors.

Additionally, the application will provide a list of customers who are most likely to churn based on the prediction model, helping you identify high-risk customers and allowing you to take proactive measures to retain them.

This user guide provides a comprehensive overview of how to set up and run the Customer Churn Prediction Application. By following these steps, you will be able to predict customer churn effectively and take data-driven actions to improve customer retention.

Application benefit:

The churn prediction application will enable our company to:

- Increase Customer Retention: By identifying at-risk customers and implementing targeted retention strategies.
- Optimize Marketing Efforts: Increase revenue by maintaining customer relations
- Enhance Customer Satisfaction: Provide personalized interventions that improve overall customer experience.
- Enable Data-Driven Decisions: Use insights from the model to make informed decisions regarding customer retention strategies.

Data Summary

- Provide the source of the raw data, how the data will be collected, or how it will be simulated.
  The source of the data will be from publicly available datasets, specifically the "Customer Churn Dataset" from Kaggle. The proposed application is going to be built with this dataset in mind, so for future use, our company can format the data collection process based on the one provided by Kaggle.

- Describe how data will be processed and managed throughout the application development life cycle: design, development, maintenance, or others.

  - **Design Phase**: in this phase, we will conduct data analysis to understand the structure and characteristics of the data. This will involve identifying key features that influence customer churn and determining the best methods for preprocessing the data.
  - **Development Phase**: The data will undergo preprocessing to ensure it is clean and ready for model training. This includes handling missing values, encoding categorical variables, scaling numerical features, and addressing any outliers. We will use industry-standard techniques to ensure data quality and integrity.
  - **Maintenance Phase**: After deployment, continuous data management will be critical. We will establish procedures for regularly updating the model with new customer data to keep it accurate and relevant. This will involve monitoring the model's performance and retraining it as necessary with fresh data.

- Justify why the data meets the needs of the project. If relevant, describe how data anomalies, e.g., outliers, incomplete data, etc., will be handled.

    - The data includes various attributes that correlate with customer behavior and churn risk. This data is well-suited for the project because it captures the essential aspects of customer interactions that are relevant to predicting churn. The Kaggle Customer Churn Dataset will be the initial testing ground to validate our methods and ensure the model's accuracy before applying it to our internal data.
    - Handling data inconsistency, such as incomplete data, is a crucial part of our preprocessing steps. We will use imputation techniques to fill in missing values and employ methods to detect and manage outliers.

- Address any ethical or legal concerns regarding the data. If there are no concerns, explain why.
    - There are no ethical or legal concerns regarding the data that will be used in the prediction application. The data to be used consists of basic customer information such as age, tenure, total spending, and payment delays. This type of data does not raise any ethical or legal issues and is commonly used in business analytics.

Implementation

- Phase 1: Data Collection and Preprocessing
    - Collect and preprocess customer data.
    - Handle missing values, outliers, and feature encoding.
    - Perform data analysis.

- Phase 2: Model Development
    - Develop the churn prediction model using machine learning techniques.
    - Train different models using customer data.
    - Find the best model from the ones tested
    - Optimize the model for accuracy and performance.

- Phase 3: UI Development
    - Develop a simple user interface for churn prediction.
    - Integrate the model with the UI, and keep it simple for user accessibility.

- Phase 4: Testing and Validation
    - Perform testing of the model
    - Validate the results with a subset of customer data.

- Phase 5: Deployment and Training
    - Deploy the model and UI.
    - Train the customer retention team on using the application.

- Phase 6: Monitoring and Maintenance
    - Monitor the model's performance.
    - Update the model with new data as necessary.
    - Provide continuous support and maintenance.

Timeline

- Provide a projected timeline, including projected start dates and end dates for each milestone (a table is not required but encouraged).

Milestone or deliverable	Duration (days)	Projected start date	Anticipated end date
Data Collection	60	6/01/2024	8/01/2024
Model Development	30	8/01/2024	9/01/2024
Model Testing	30	9/01/2024	10/01/2024
Team Training	14	10/01/2024	10/15/2024

Evaluation Plan

- Describe the verification method(s) to be used at each stage of development.
    - During development: Regular code reviews and testing will be done throughout the data collection, model development, and model testing stages
    - Post-development: User acceptance testing will be performed to ensure the application meets business requirements and accessibility for users.

- Describe the validation method to be used upon completion of the project.
    - The validation method to be used on the application upon completion, will evaluate the models' performance using different sets of customer data. Metrics such as accuracy and precision will be used to judge the application predictions.

Resources and Costs

- Hardware and Software Costs
    - Server Infrastructure: $5,000
    - Software Licenses: $3,000
    - Development Tools: $2,000

- Labor cost
    - Data Scientist: 4 months at $10,000 monthly = $40,000
    - Developers: 4 months at $8,000 monthly = $32,000
    - Project Manager: 4 months at $12,000 monthly = $48,000

- Itemize estimated environment costs of the application, e.g., deployment, hosting, maintenance, etc.
    - Deployment: $5000
    - Hosting: 4 months at $1000 monthly = $12,000 yearly
    - Maintenance: 4 months at $1000 monthly = $12,000 yearly

- Total Estimated Costs
    - Initial setup: $10,000
    - Labor Cost: $120,000
    - Environment Costs: $24,000 yearly
    - Total = $154,000

Part D: Post-implementation Report

Solution Summary

- Summarize the problem and solution.

  Our company, like many other companies, faced a significant challenge presented by customer churn. This negatively impacted revenue and growth. Customer churn occurs when clients or subscribers discontinue their service and put an end to their relationship with a company. The solution I presented, provides the key to financial growth for our company. The solution is to predict which customers are at high risk of churning. With the ability to predict which customers are likely to churn, proactive action to retain the customer can be taken.

This prediction model uses machine learning techniques. The model with the help of algorithms analyzes various customer data points such as age, tenure, usage frequency, payment delays, subscription type, and many more to predict the likelihood of each customer churning. By identifying patterns and trends, the model provides insights that enable our company retention team to focus on those specific customers.

- Describe how the application provides a solution to the problem from parts A and B.
  - Data Collection and Preprocessing: We obtained the testing data from a publicly available dataset from Kaggle specifically the one titled "Customer Churn Dataset". The preprocessing phase was concerned with how the variables with missing values were managed. The categorical data were encoded while the numerical features were scaled. The data was then ready for model tuning and evaluation.
  - Model Development and Training: An analysis of customer churn was made and we adopted a machine-learning approach of training a model for churn predictions. The model was trained using customer data and was optimized for accuracy. We used various algorithms, consisting of a Decision Tree, Naïve Bayes, and Support Vector Machine (SVM), to identify the best-performing model.
  - Interactive Dashboard: An interactive dashboard was created to allow users to input customer data and receive churn predictions. This dashboard provides a user-friendly interface for the customer retention team to analyze and act on the predictions.
  - Evaluation and Validation: The model's performance was evaluated using different metrics such as accuracy and precision. We conducted user acceptance testing to ensure the project requirements were met by the application.
  - User Guide and Technical Documentation: To ensure that the users of the churn prediction model and the churn dashboard could be able to use them

appropriately, detailed documentation was prepared. This consisted of a User guide which explained the usage of the model.

Data Summary

The application was built around a publicly available dataset that was obtained from Kaggle. This dataset provided insightful customer information. The Kaggle dataset was chosen for its detailed attributes and high quality, which are important for building an accurate customer churn prediction model.

The data went through several stages of processing and management throughout the application development life cycle. During the design phase, we defined the data requirements based on the Kaggle dataset structure. The initial dataset was reviewed to understand its structure, attributes, and any potential issues such as missing values or categorical variables.

In the development phase, the raw data was preprocessed to ensure it was clean and suitable for training the machine learning models. This included handling missing values, encoding categorical variables, and scaling numerical features. The processed data was then split into training and testing sets to train and validate the models effectively with a 75/25 split. It's also important to mention that the size of the data used was scaled down to a small portion of the original data. Although it would have been better to use more data to train the application, a smaller portion allowed us to better visualize the data.

For our company's use, it is important to have a dataset with the same format as the Kaggle dataset since the application was built with that dataset in mind. When putting together the customer data from our company's customer database, the same customer attributes should be collected as those in the Kaggle dataset. Then by using our company's dataset that follows this format, we can effectively apply the customer churn prediction model to predict customer churn, ultimately enhancing customer retention and reducing costs.

Machine Learning

Method 1: Decision Tree Classifier

What

"A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical tree structure, which consists of a root node, branches, internal nodes, and leaf nodes.", (IBM, 2021). It works by splitting the dataset into subsets based on the value of input features, creating a tree-like structure where each node represents a feature, each branch represents a decision rule, and each leaf node represents an outcome. In this project, the Decision Tree Classifier was used to predict whether a customer would churn based on multiple different features of data.

How

The Decision Tree model was developed by first preprocessing the data to ensure it was clean and suitable for training. This involved handling missing values, encoding categorical variables, and scaling numerical features. The preprocessed data was then split into training and testing sets. The Decision Tree algorithm was applied to the training data, where it learned to make predictions by creating decision rules based on the input features.

To understand how the Decision Tree model was built, we have to first go through how the data was prepared given that it was used for training the model. This includes steps on dealing with missing values, transforming categorical attributes to numerical, and feature scaling numerical attributes. The preprocessed data was partitioned into the training and testing datasets. The Decision Tree algorithm was applied to the training data, where it learned to make predictions by creating decision rules based on the input features. After training the model it was able to make 90% accurate predictions on the test data.

Why

The Decision Tree Classifier was chosen for its simplicity. It provides clear decision rules that can be easily understood and visualized, making it an excellent choice for explaining the model's predictions. Additionally, Decision Trees can handle both numerical and categorical data, which is perfect given the nature of the customer data. The development process was straightforward, and the model outperformed the other models in terms of accuracy and precision, justifying its use in predicting customer churn.

Method 2: Naïve Bayes Classifier

What

“The Naïve Bayes classifier is a supervised machine learning algorithm that is used for classification tasks such as text classification. They use principles of probability to perform classification tasks. “, (IBM Naïve Bayes, 2021). The Naïve Bayes Classifier is a probabilistic machine learning algorithm based on Bayes' theorem. It assumes independence between the features given the class label. It is particularly useful for classification tasks and is known for its simplicity and effectiveness, especially with large datasets.

How

The Naïve Bayes model was developed the same way as the Decision Tree. Once data was preprocessed, split into training and testing datasets, and feature scaled the training dataset was applied to Naïve Bates mode. Once trained, it was able to make predictions when the test data was applied. Although not as accurate as the Decision tree model with an accuracy score of 87%.

Why

The Naïve Bayes Classifier was selected for use due to its fast and simple computational capabilities. For large datasets, it is exceptionally adaptive, and it offers fast predictions, which are critical in real-time models. Naïve Bayes might not be the most sophisticated model out there, but the algorithm is effective and serves as a good starting point for any complex algorithm. This made it even more appealing for this project since the model could handle categorical data directly without the need for much preprocessing.

Method 3: Support Vector Machine (SVM)

What

“SVMs is a powerful machine learning algorithm, it can be used for a variety of tasks, such as text classification, image classification, spam detection, handwriting identification, gene expression analysis, face detection, and anomaly detection.”, (GeeksforGeeks, 2023). It works by finding the hyperplane that best separates the data into different classes. The SVM algorithm maximizes the margin between the hyperplane and the nearest data points from each class, known as support vectors.

How

The SVM model was also developed in the same way as the other two models. Once data was preprocessed, split into training and testing datasets, and feature scaled the training dataset was applied to SVM mode. Once trained, it was able to make predictions when the test data was applied. Although not as accurate as the Decision tree or Naïve Bayes models with an accuracy score of 85%.

Why

SVM was chosen because it is less affected by noise and performs well in higher dimensional space. It is best in situations that can have a measurable separation between classes. An added advantage of SVM is that it is effective in the case of both linear as well as non-linear decision boundaries. The use of SVM was made based on the decision that it was among the most accurate when working with complex data of customer churn.

Justification of Method Selection and Development

Each of the above methods was chosen due to their capabilities for handling the customer churn prediction problem. The Decision Tree Classifier was used as a result of its ability in interpretability and capacity to handle mixed inputs. Naïve Bayes was selected for its simplicity and efficiency, providing a strong baseline model. SVM was used primarily based on its stability and ability to work with data spaces of high dimensions.

Validation

<u>Decision Tree Classifier</u>

Validation Method

The validation method used on the Decision Tree Classifier was a 75/25 train-test split. This method involves splitting the dataset into two parts: 75% of the data was used for training the model, and the remaining 25% was used for testing its performance.

Results

The Decision Tree Classifier returned with an accuracy of 90% on the test set. The confusion matrix and classification report for the test set indicated high precision, confirming the model's effectiveness in predicting customer churn.

Naïve Bayes Classifier

Validation Method

The Naïve Bayes Classifier was also validated using the same 75/25 train-test split. This consistent approach allowed for a fair comparison between models.

Results

The Naïve Bayes Classifier achieved an accuracy of 87% on the test set. The confusion matrix and classification report for the test set showed good performance, though slightly lower than the Decision Tree model.

Support Vector Machine (SVM)

Validation Method

Like the other models, the Support Vector Machine (SVM) model was validated using the 75/25 train-test split. This method provided a consistent way of evaluating the performance of all models.

Results

The SVM model acquired an accuracy of 85% on the test set. The confusion matrix and classification report indicated that while the SVM model performed well, it was slightly less accurate than the Decision Tree and Naïve Bayes models.

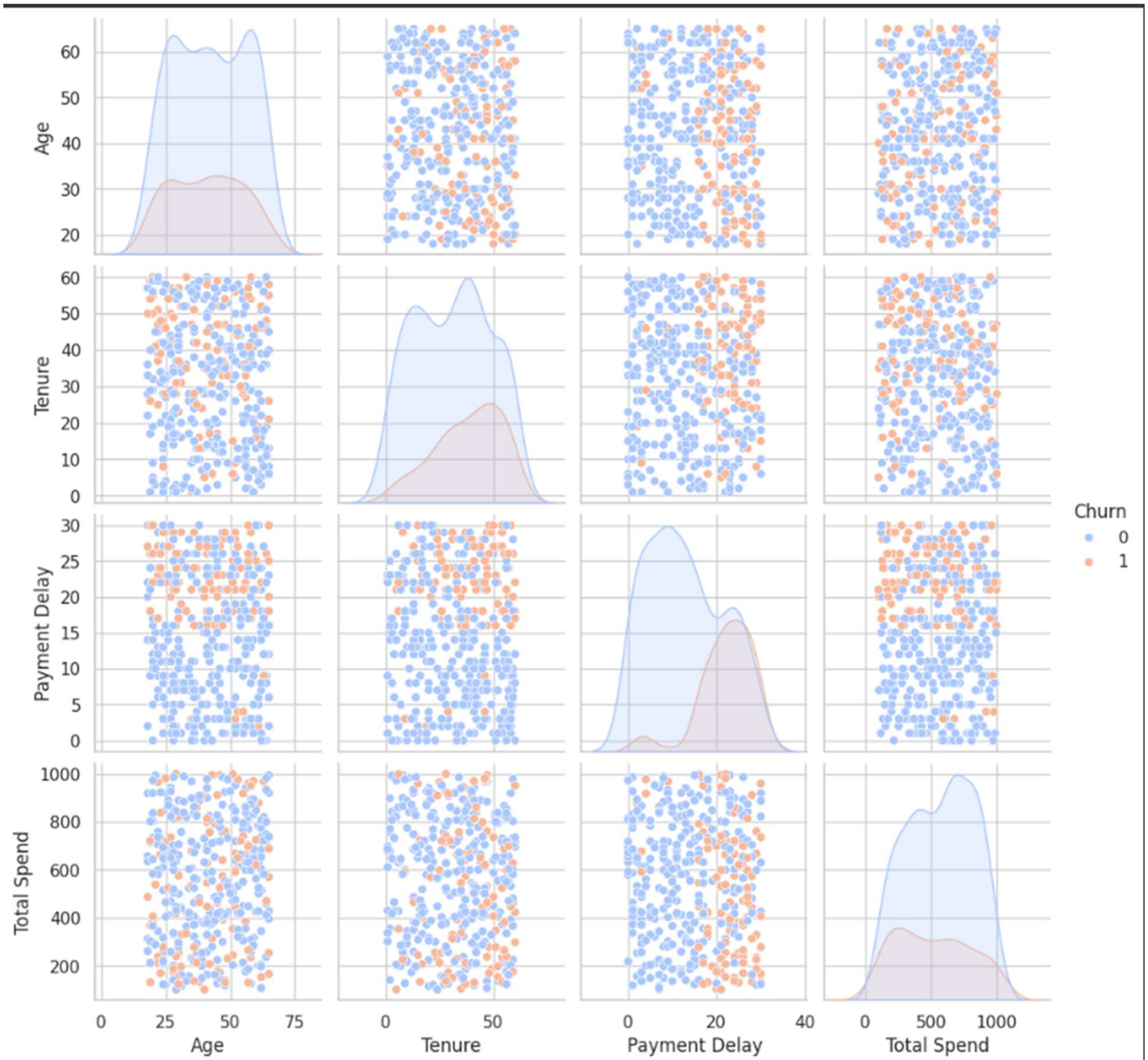
Future Plan for Validation

For future validation, more elaborate data sets and more validation methods such as k-fold validation could be used to validate the models. The model was only tested using 400 rows of data only. This might sound significant, yet in real-life examples, businesses deal with thousands of records. The more data used increases the accuracy and reliability of the predictions. Instead of training the model on just 300 rows, imagine training it on 75% of a 40,000-row dataset, that would mean using 30,000 rows for training. This larger training set provides the algorithm with more information and context, leading to more precise predictions. For simplicity, a small portion of the actual data from Kaggle was used in this project to guarantee the readability of visuals like scatter plots. However, scaling up the data size would enhance the model's performance and accuracy.

Visualizations

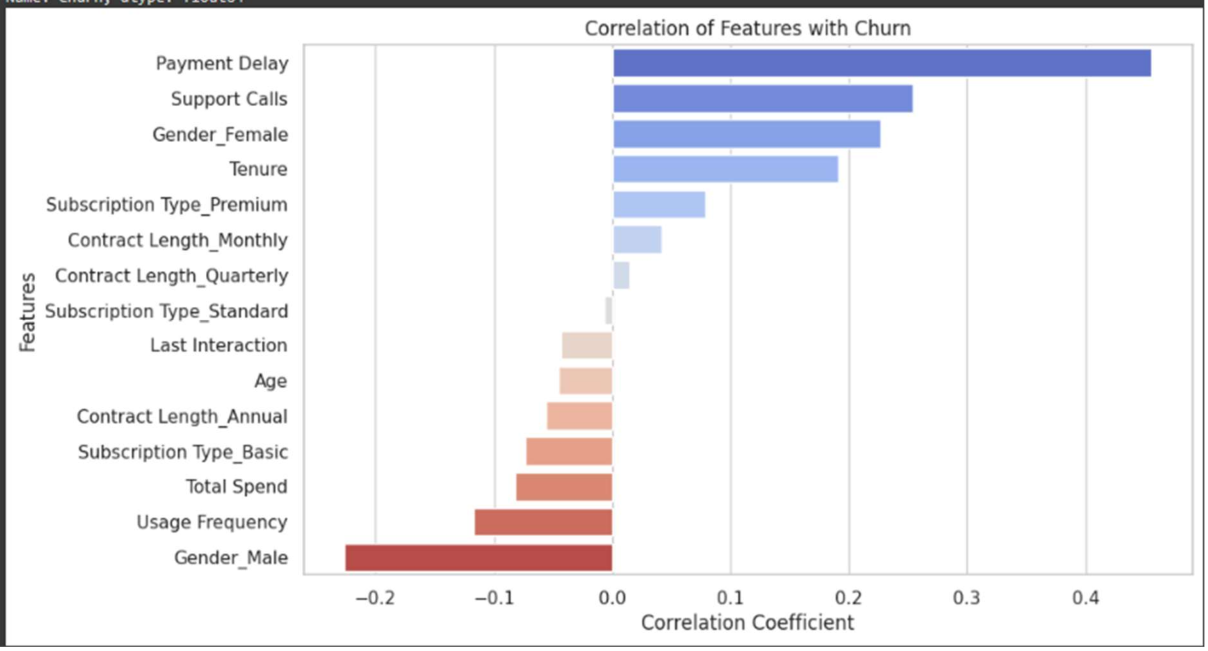
These are some visuals that have been obtained from the application. These visuals give us a better understanding of our data.

Visual #1: In picture number 1 a pair plot is used. This pair plot shows the correlation of multiple different features of the data.

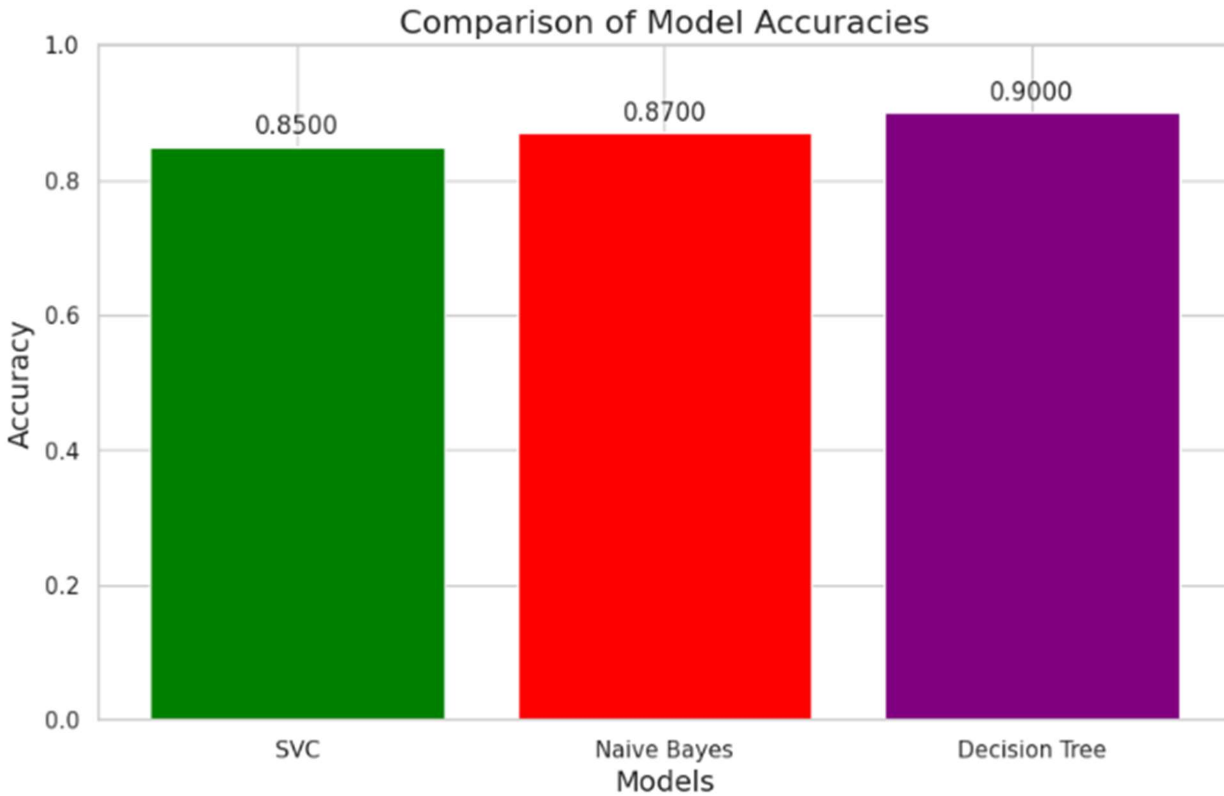


Visual #2: In picture 2, a correlation plot shows us which features have a major play in predicting if a customer will churn. The closer the feature is to 1 the more it affects the chances of customer churn.

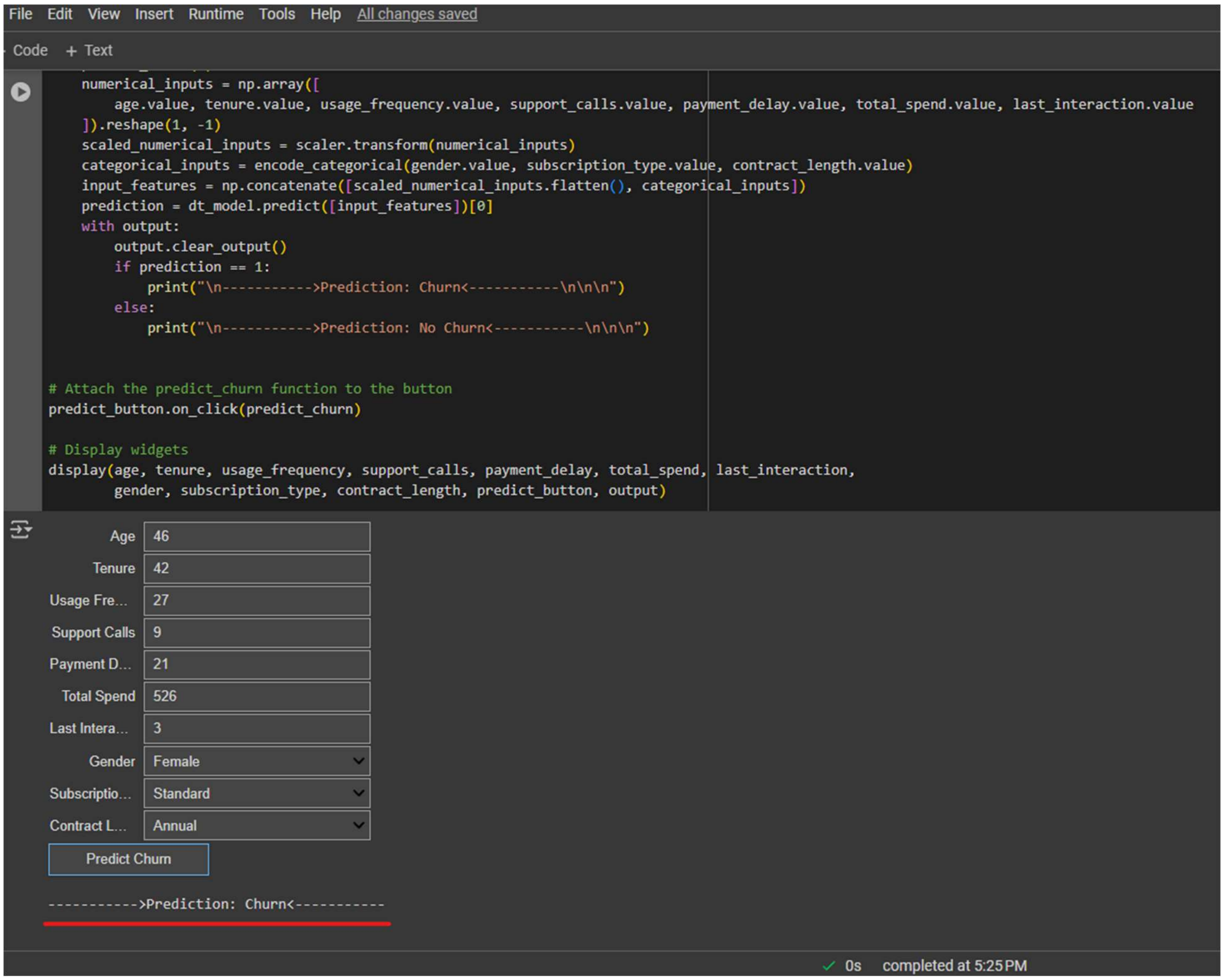
```
Correlation of features with Churn:
 Payment Delay                0.455768
Support Calls                 0.253609
Gender_Female                 0.226728
Tenure                        0.190790
Subscription Type_Premium     0.078866
Contract Length_Monthly       0.042188
Contract Length_Quarterly     0.015215
Subscription Type_Standard   -0.006584
Last Interaction             -0.043380
Age                          -0.045531
Contract Length_Annual       -0.055724
Subscription Type_Basic      -0.073597
Total Spend                  -0.081990
Usage Frequency              -0.117131
Gender_Male                  -0.226728
Name: Churn, dtype: float64
```



Visual #3: P 3 shows us a bar graph that compares the accuracy scores of the three different algorithms. This graph helped us decide which algorithm is the most fit for this application.



Visual #4: This picture shows the User Input and the prediction the model makes. Whether the customer will “Churn” or “No Churn” using the application.



User Guide

Step-by-Step Instructions to Execute and Use the Application

Using Jupyter Notebook

1. Download and Install Necessary Software or Libraries

1.1 Install Python and Jupyter Notebook:
- If you don't have Python installed, download it from python.org.
- Install Jupyter Notebook by running the following command in your terminal or command prompt:
`pip install notebook`

1.2 Install Necessary Libraries:
- Run the command:
`pip install pandas numpy matplotlib seaborn plotly scikit-learn ipywidgets termcolor`

This command should install all the necessary libraries to run the application

1.3 Open a new Notebook:
- Run the command 'Jupyter notebook' this will take you to a new notebook in Jupyter.

2. Upload the Application Code and Dataset

2.1 Upload the Code:
- Download the application code file and upload it to your Jupyter Notebook environment.

2.2 Upload the Dataset:
- Download the provided dataset CSV file and upload it to the same environment in your notebook.

3. Open and Configure the Application Code

3.1 Read the Dataset:
- Locate the line where the dataset file is being read. It should be somewhere at the beginning of the code and it looks like this:
`df = pd.read_csv('your_dataset_file.csv')`

- Ensure the file name matches the name of your uploaded dataset file. For example:
`df = pd.read_csv('customer_churn.csv')`

4. Run the Application Code

4.1 Execute All Cells:

- In Jupyter Notebook, use the "Run All" option to execute all cells in the notebook. This will preprocess the data, train the models, and generate predictions and visualizations.
- or you can run each cell, but make sure to run them sequentially or there will be errors

5. Interact with the Application

5.1 Explore the Visualizations:

- The code will generate various visualizations, such as scatter plots, bar charts, and ROC curves, which will provide you insight based on the data.

5.2 User Input

- Towards the end of the code, there will be a section where you can input data about a customer and it will give you predictions on churn or no churn.

Using Google Colab

If you are using Google Colab, you can simplify the setup process since all the necessary libraries are pre-installed. Follow these steps:

1. Open Google Colab in your web browser.

2. Upload the application code file and dataset to your Colab environment.

3. Ensure the dataset file name in the code matches your uploaded file name.

4. Run all cells in the notebook to execute the code and generate the necessary visualizations and predictions.

Example Usage

The application allows you to input a customer dataset and it returns predictions on whether individual customers are likely to churn. Multiple visualizations are provided with the application to help you understand patterns and trends within the data used. Also, the application provides an interactive interface that allows you to enter individual customer data and it predicts if they will churn. The insight provided by the application helps you make informed decisions about customers and helps focus on those likely to churn.

Reference Page

1. *What is a decision tree?*. IBM. (2021, November 2). https://www.ibm.com/topics/decision-trees#:~:text=A%20decision%20tree%20is%20a,internal%20nodes%20and%20leaf%20nodes.

2. *What are naïve Bayes classifiers?*. IBM. (2021a, October 6). https://www.ibm.com/topics/naive-bayes#:~:text=The%20Na%C3%AFve%20Bayes%20classifier%20is,probability%20to%20perform%20classification%20tasks.

3. *Support Vector Machine (SVM) algorithm*. GeeksforGeeks. (2023, June 10). https://www.geeksforgeeks.org/support-vector-machine-algorithm/