Data Science Intern at Data Glacier

Week 13: Deliverables

Name: Tejeswar Reddy Nalijeni

University: University of Cincinnati

Email: nalijety@mail.uc.edu or tejeswarreddyn2808@gmail.com

Country: United States

Specialization: Data Science

Batch Code: LISUM35

Date: Sep 30th, 2024

Submitted to: Data Glacier

Table of Contents:

1. Project Plan

Weeks	Date	Plan
Weeks 07	Aug 19, 2024	Project Preparation, Data Intake Report
Weeks 08	Aug 26, 2024	Data processing
Weeks 09	Sept 2, 2024	Data Processing (Advanced)
Weeks 10	Sept 9, 2024	Data Analysis, EDA
Weeks 11	Sept 16, 2024	Build Model Preparation
Weeks 12	Sept 23, 2024	Explore Different Model
Weeks 13	Sept 30, 2024	Presentation for data result & Model Evaluation, Code

2. Problem Statement and Understanding

2.1. Problem Description

The data is related to a company where they have various subscription plans based on tenure and provide support calls to customers if they have any issues or if there is any churn. Our goal is to predict the reasons for customer churn (variable y) and analyze the data to increase customer subscriptions.

2.2. Business Understanding

The main goal of this project is to predict the reasons behind customer churn in subscriptions using recorded data. This involves optimizing marketing efforts, improving customer engagement strategies, and ultimately boosting subscription numbers. By utilizing historical data for binary classification, the project seeks to accurately identify potential subscribers, understand the causes of churn, and increase customer retention through targeted and informed promotions.

3. Data Intake Report

Click here for **Data Intake Report**.

4. Data Understanding

4.1. Columns

1. Customer ID: A unique identifier for each customer.
2. Age: The age of the customer.
3. Gender: The gender of the customer.

4. Tenure: The length of time the customer has been with the service.
5. Usage Frequency: How often the customer uses the service.
6. Support Calls: The number of support calls made by the customer.
7. Payment Delay: The number of times the customer has delayed payment.
8. Subscription Type: The type of subscription the customer has.
9. Contract Length: The duration of the customer's contract.
10. Total Spend: The total amount of money the customer has spent.
11. Last Interaction: The time since the customer's last interaction with the service.
12. Churn: Indicates whether the customer has churned (likely a binary column with values like 0 for 'No' and 1 for 'Yes').

4.2. Information

As we haven't started the analysis yet, we can't determine if the data contains outliers, missing values, etc. However, we can outline a strategy for handling missing values and outliers.

Handling Missing Values:

1. If there are many missing values:
   - For categorical variables, replace missing values with the most common (mode) value.
   - For continuous variables, replace missing values with the median.
2. If the missing values are few:
   - Consider simply deleting those rows.

   So, as my data have only one missing row, I considered to delete that row which does not impact the result.

Handling Outliers:

This will be assessed on a case-by-case basis, depending on the dataset. We'll address this as we proceed with the analysis. “Age” appears normal but skewed to the left. “Total spend” appears normal bust skewed to right. Almost all the graphs except “Total spend` have strong Negative/left-skewed.

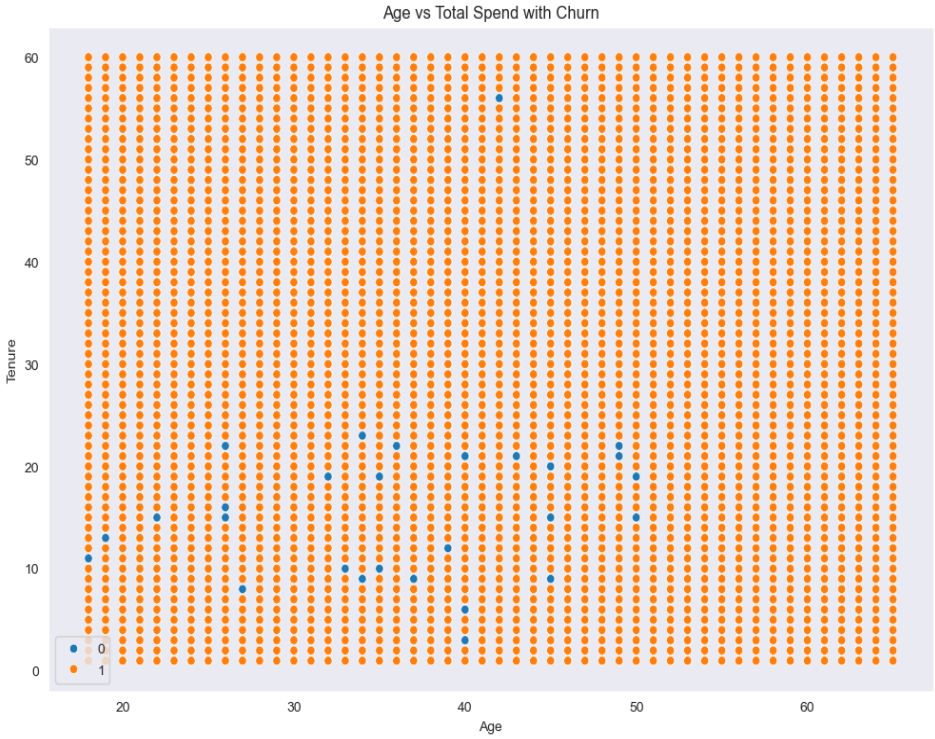
Campaign-Specific Data: The target variable for our predictive model is "y," indicating whether the customer churns or not.

5. Exploratory Data Analysis (EDA)

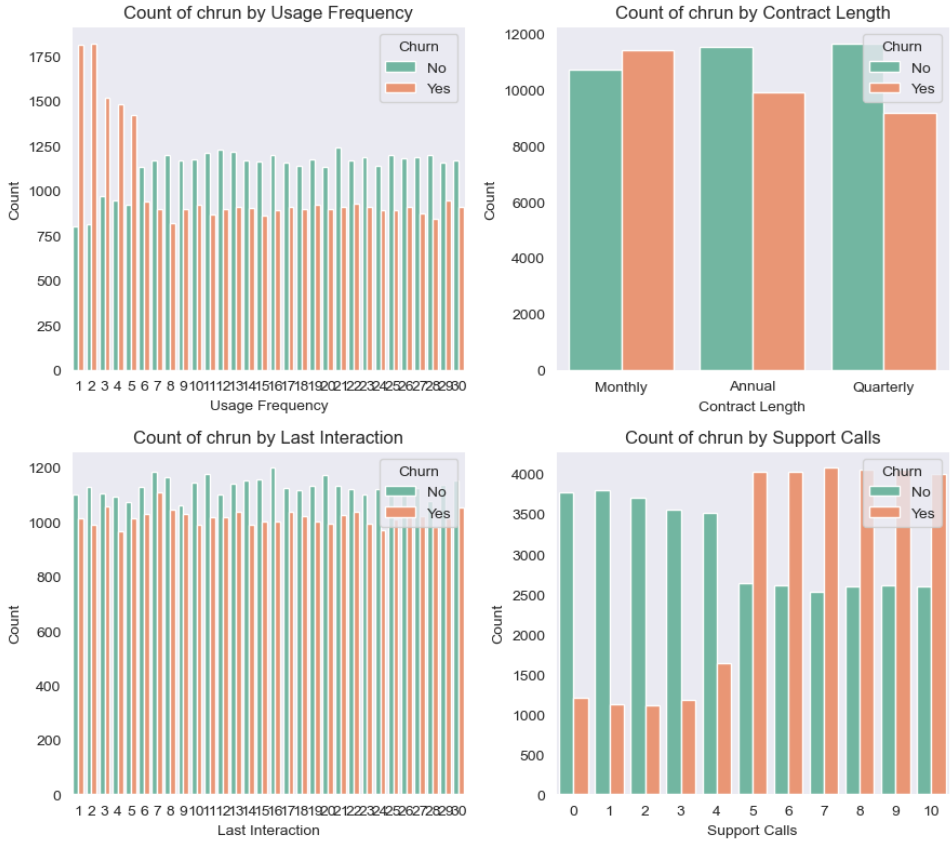
In this week, I conducted an in-depth analysis of the customer churn dataset, focusing on customer demographics and other relevant attributes. The analysis included a heatmap to explore potential correlations between variables. However, no clear relationships between the variables have been identified so far that would allow me to make any strong recommendations. That said, I anticipate the upcoming modeling phase will highlight the main factors influencing customer churn, which will provide actionable insights. I should be able to offer final recommendations after that.

- The scatter plot does not show a clear trend between age and total spend in relation to churn. However, it indicates that customers aged between 20 and 50 with a total spend over 500 are more likely to churn. In contrast, those with lower total spend appear less prone to churning. This suggests a potential correlation between higher spending and churn within this age group. Further analysis may be needed to confirm

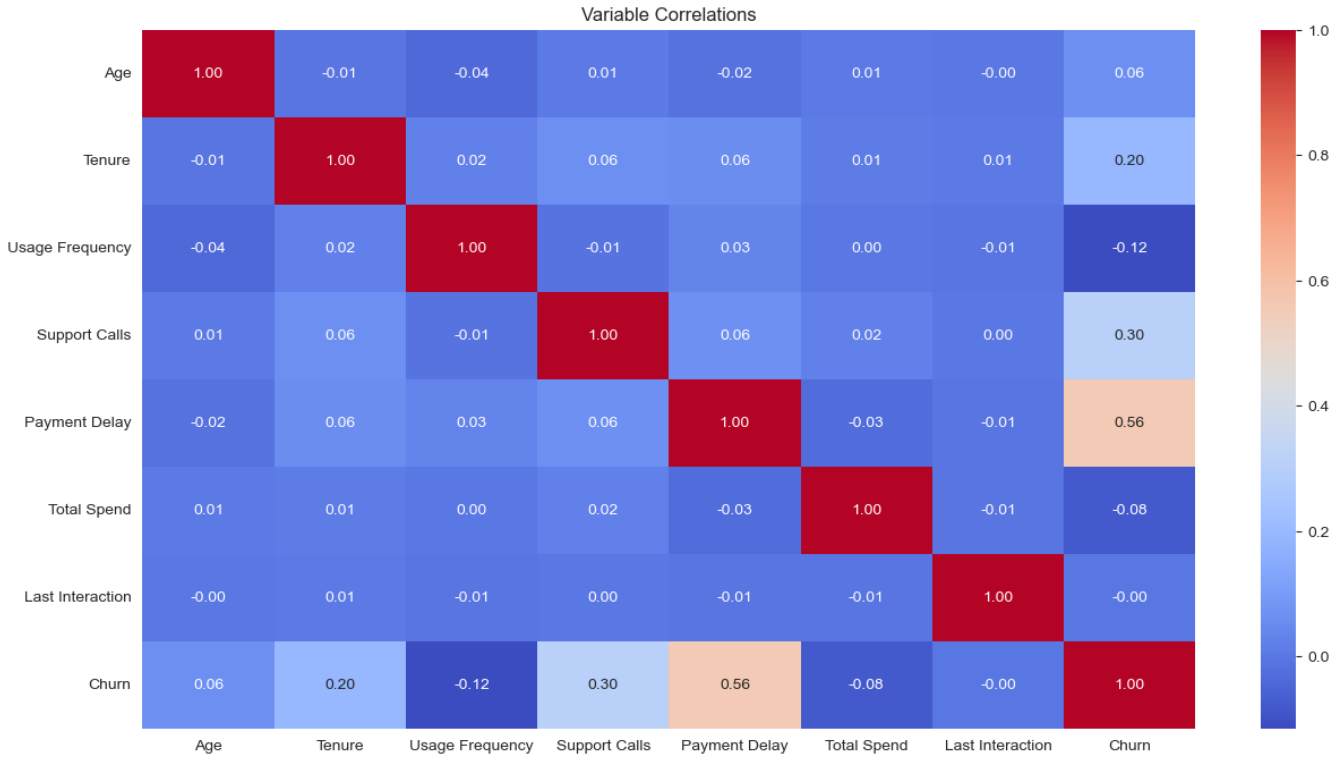
this pattern.



- The bar chart shows that while there are differences in subscription types acrossUsage Frequency, Contract Length, Last Interaction, Support Calls but not all have a great information.



- The below is the correlation heatmap for numerical variables.



6. Explore Distinct Model

The code begins by checking for any missing values (NaNs) in the training and test datasets. To handle this, it fills in the missing values using the average of the available data and removes any rows with NaNs afterward. Once the data is cleaned, several machine learning models are trained, including Logistic Regression, Gradient Boosting, Random Forest, Decision Tree, SVM, KNN, XGBoost, and LightGBM.

After training the models, they are tested on the test dataset, and their accuracy is calculated. The results show that ensemble models like XGBoost, LightGBM, Gradient Boosting, and Random Forest perform exceptionally well, achieving near-perfect accuracy. In contrast, traditional models like Logistic Regression and KNN show lower accuracy but still perform decently. Overall, the ensemble methods significantly outperform the others in terms of accuracy.

7. Model Evaluation

The evaluation of multiple machine learning models reveals varying performance levels across metrics like accuracy, precision, recall, F1 score, and AUC-ROC. Logistic Regression performs reasonably well with 82.79% accuracy and a solid balance between precision (81.52%) and recall, along with a good

AUC-ROC score of 90.51%. Gradient Boosting, Random Forest, XGBoost, and LGBM stand out as the top performers, achieving near-perfect accuracy (99.69%-99.99%) and excellent precision and recall scores. Notably, XGBoost and LGBM reached a perfect AUC-ROC of 1.0000, demonstrating their superiority in classification tasks. Decision Tree Classifier also performed exceptionally well, achieving 99.94% accuracy with high consistency across all metrics. In contrast, SVM and KNN, while still effective, show relatively lower performance, with SVM achieving 93.30% accuracy and KNN at 90.86%. While traditional models like Logistic Regression and KNN perform decently, ensemble methods like Random Forest, XGBoost, and LGBM clearly outperform them, excelling in both accuracy and precision.

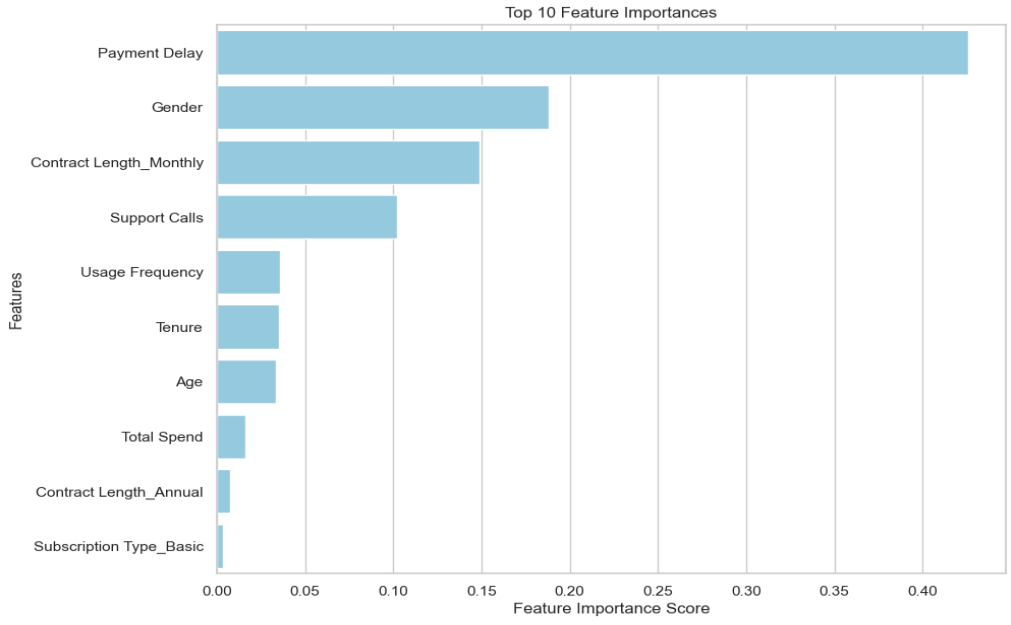
8. Model Comparison

After evaluating all eight models on the dataset, we can conclude that the XGB Classifier stands out as the top performer, achieving the highest accuracy of 0.9999. It slightly outperforms the LGBM Classifier (0.9998) and the Random Forest Classifier (0.9992) in terms of accuracy. The Decision Tree Classifier also performs exceptionally well, with an accuracy of 0.9994, but falls just behind the XGB Classifier. Notably, the XGB Classifier not only leads in accuracy but also achieves a perfect AUC-ROC score of 1.0000, demonstrating its superior ability to distinguish between classes. Based on these metrics, the XGB Classifier is recommended as the most reliable and effective model among those tested.

9. Best Feature

The feature importance analysis from the XGBoost model reveals that Payment Delay is the most significant predictor, contributing the highest importance score of 0.4258. Other influential features include Gender (0.1883) and Contract Length (Monthly) (0.1491), followed by Support Calls (0.1020) and Usage Frequency (0.0358). Tenure and Age also play moderate roles, while Total Spend, Contract Length (Annual), and Subscription Type (Basic) show comparatively lower importance. These top 10 features highlight the key factors that the model relies on to make predictions, with payment-related factors being the most critical.

- **Payment Delay:** Longer payment delays are strongly associated with higher churn rates, indicating that customers who delay payments are more likely to leave.
- **Support Calls:** A higher number of support calls is weakly associated with churn, suggesting that customers needing frequent assistance may be more prone to churn.
- **Tenure:** There is a weak positive correlation, indicating that customers with longer tenures are slightly more likely to churn, though the relationship is not very strong.



10.Recommendation

- **Focus on Payment Delay**: Allocate more resources to resolving payment delays, as models like XGB Classifier, Random Forest, and LGBM consistently show that this is the most influential factor for predicting churn. Use predictive insights to proactively identify customers at risk of delayed payments and implement personalized reminders or flexible payment options to reduce churn.
- **Address Support Issues**: Models such as Gradient Boosting and SVM highlight the correlation between support calls and churn. High support call volumes often indicate customer dissatisfaction. Enhance support services and provide faster issue resolution to minimize churn among customers needing frequent assistance.
- **Refine Approach Using Model Insights**: Leverage the high accuracy and interpretability of models like Random Forest and Decision Tree to refine your churn prevention strategies. Use feature importance analysis to identify high-risk customers early and develop targeted retention campaigns. These models can help focus marketing efforts on the most critical factors, such as payment behaviors and support experiences, ensuring better resource allocation and improving overall retention.