PLATA

Agenda

- INITIAL ANALYSIS
- EDA
- OUTLIERS AND MISSING VALUES
- APPROACHES TO MODELLING
- INITIAL INVESTIGATION
- TIME SERIES MODELLING
- ADDITIONAL APPROACH
- LIMITATIONS

INITIAL ANALYSIS

Distribution by Date:

Applications (APPLICATION_DATE): Data available for 74 days. Daily applications range from 9 to 593

Agreements (AGREEMENT_DATE): Data available for 71 days. Daily agreements range from 3 to 345.

Card Uses (UTILIZATION_DATE): Data available for 71 days. Daily card uses range from 2 to 252

Stage Transitions:

Agreement Conversion Rate: 49.04% (about half of applications lead to agreements)

Card Utilization Rate: 80.74% (approximately 81% of agreements result in card use)

Number of Signed Agreements Over Time

Number of Card Uses Over Time

INITIAL ANALYSIS

Signed Agreements Leading to Card Use

Among the agreements that led to card utilization, the average time from application to agreement is approximately 7.5 days, with a median of 5 days. Once the agreement is signed, the average time to card utilization is about 4.3 days, and the median is 2 days. The utilization time ranges from the same day to a maximum of 61 days, with a standard deviation of 6.1 days, indicating some variability in how quickly cards are used after the agreement is finalized.

Signed Agreements with No Card Use

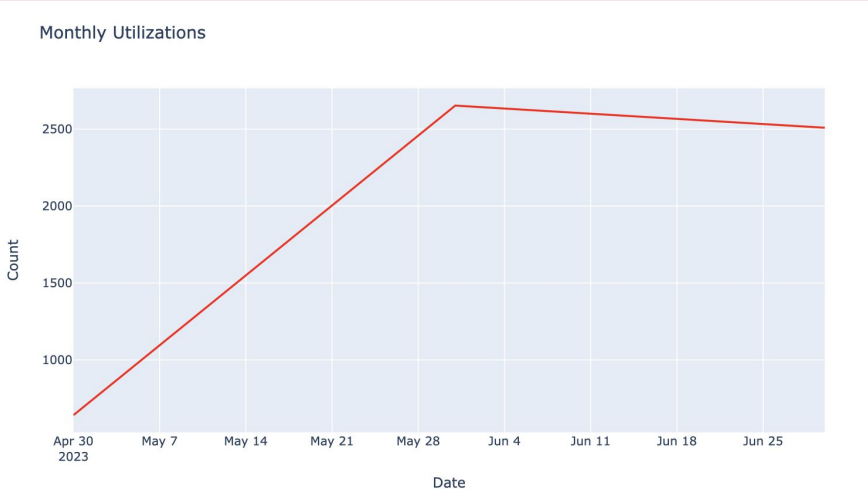
For agreements that did not result in card use, the average duration from application to agreement is 9.5 days, with a median of 6 days. The agreement was sometimes signed on the same day as the application, with a maximum delay of up to 60 days. The standard deviation is 9.7 days, reflecting substantial variability in the time taken to finalize these agreements. Notably, about 15.6% of all signed agreements did not lead to card utilization.

Signed Agreements Leading to Card Use

Signed Agreements with No Card Use

UTILISATION

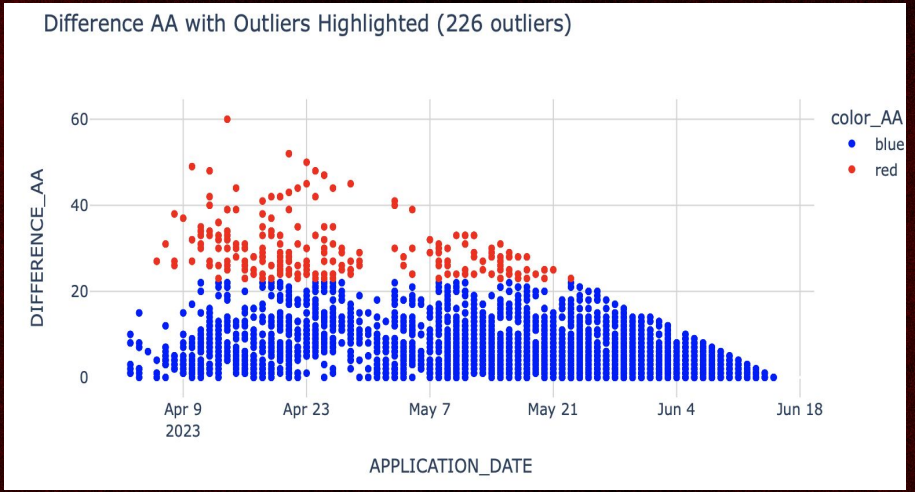
A trend where the daily, weekly, and monthly utilization metrics initially increase over time and then begin to decrease can be observed in the data

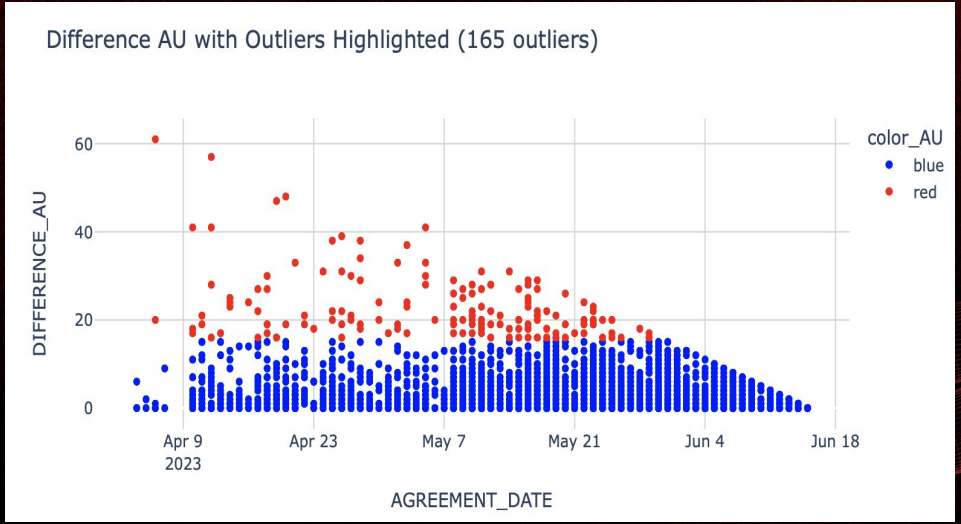


Initial Analysis

The dataset contains missing values primarily in the UTILIZATION_DATE column, indicating instances where the card was not utilized after the agreement was signed.

The absence of a UTILIZATION_DATE suggests that despite an agreement being signed, the user did not proceed to use the card. This could indicate potential issues with the user experience, the product's appeal, or external factors affecting utilization.





'DIFFERENCE_AA' (Time between Application and Agreement): There are 226 outliers, representing unusual cases with either very short or very long times between application and agreement.

'DIFFERENCE_AU' (Time between Agreement and Utilization): There are 165 outliers, indicating significant variability in the time taken for card utilization after the agreement.

OUTLIERS AND MISSING VALUES

IQR AND OUTLIERS

Interquartile Range (IQR) helps to identify outliers by measuring the spread of the middle 50% of the data. Values that fall below Q1−1.5×IQR Q1−1.5×IQR or above Q3+1.5×IQR Q3+1.5×IQR are typically considered outliers.

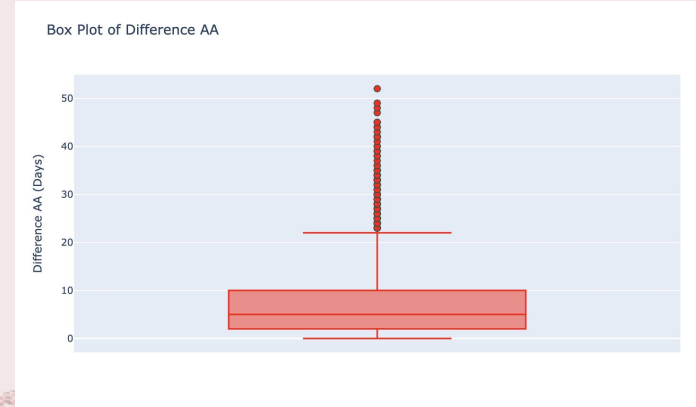
OUTLIERS:

DIFFERENCE_AA: 300 outliers

DIFFERENCE_AU: 689 outliers

MISSING VALUES:

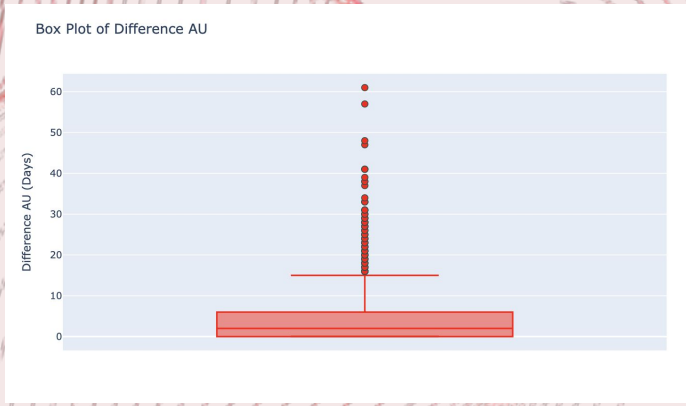
Missing Values were Replaced with 0



DATA DIFFERENCE

DIFFERENCE_AA: Days between AGREEMENT_DATE and APPLICATION_DATE.

DIFFERENCE_AU: Days between UTILIZATION_DATE and AGREEMENT_DATE.



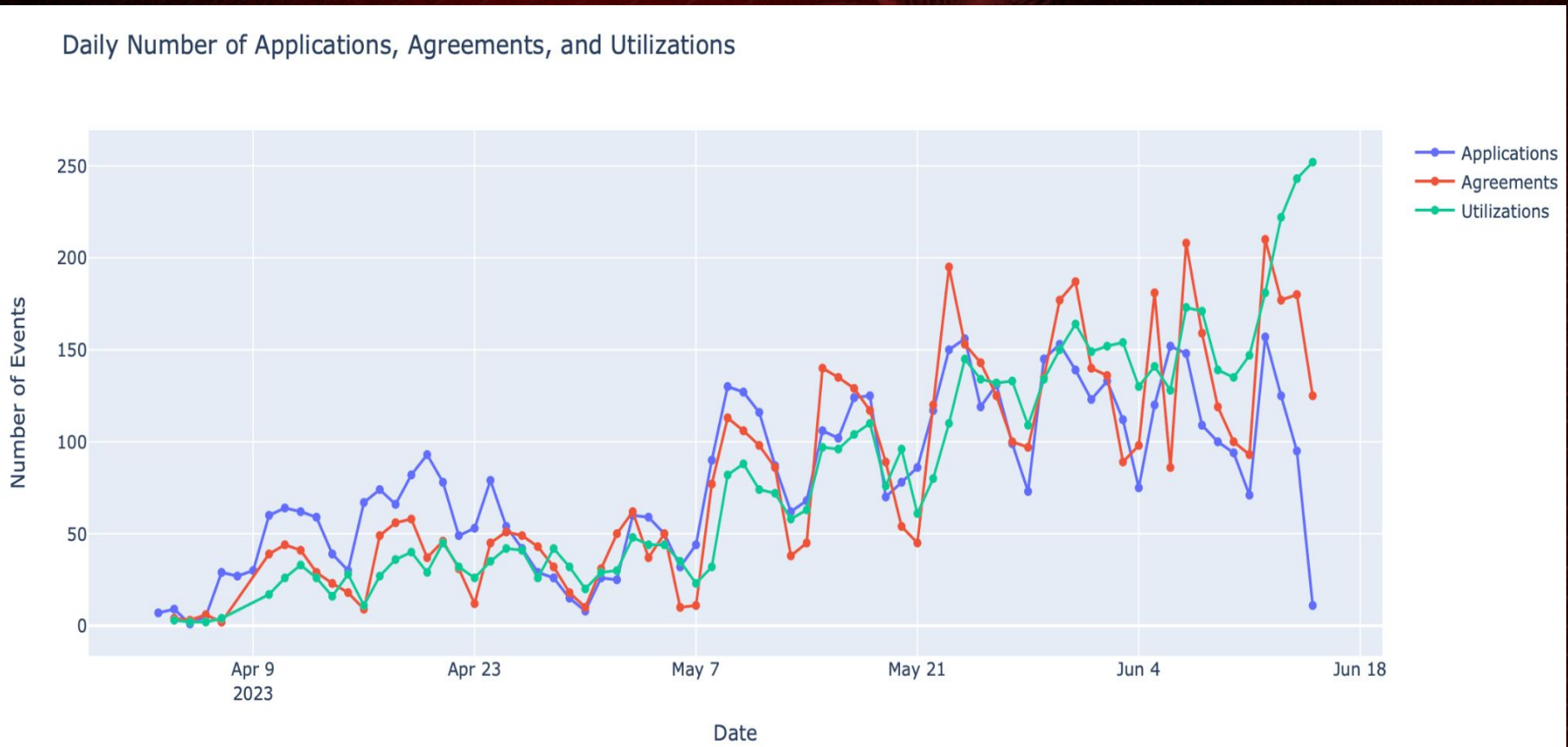
REMOVED VALUES

Eliminated rows with negative differences.

Filtered DIFFERENCE_AA values greater than 45 days.

Filtered DIFFERENCE_AU values greater than 41 days.

Final Shape 2791 rows × 8 columns



Despite the decrease in applications and agreements, there's a significant rise in utilization, meaning that those who have cards are using them more frequently.

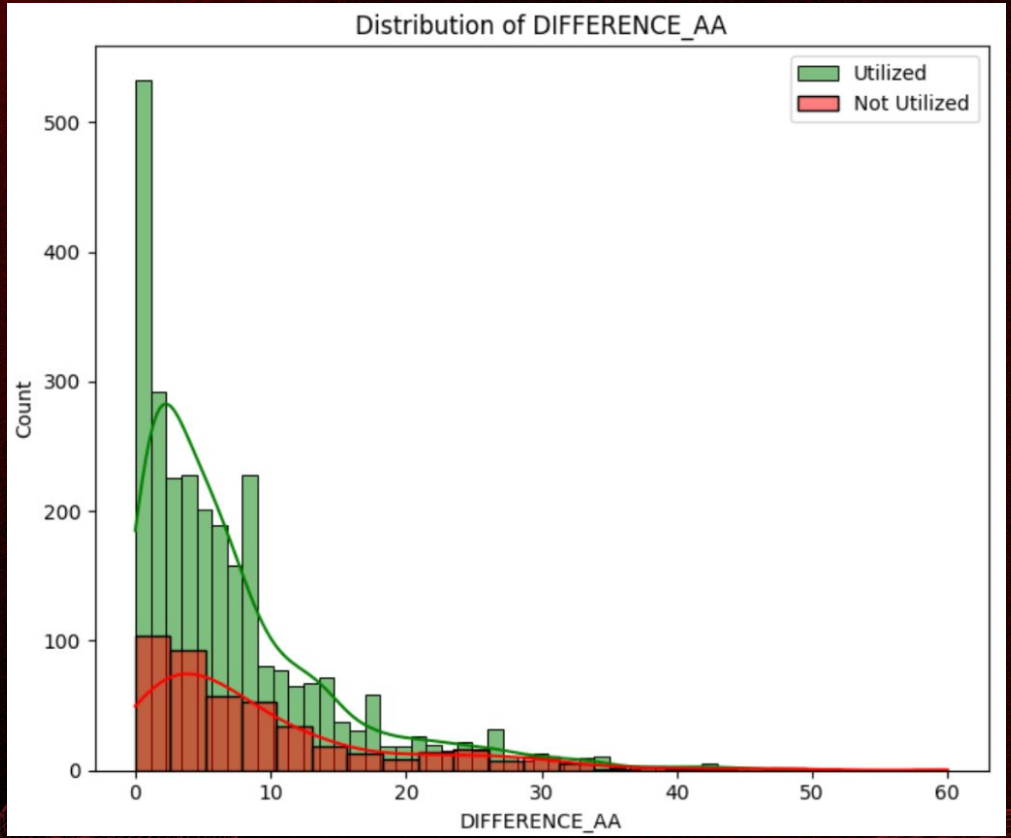
Gap Analysis:

The majority of utilized cards had no significant delay between agreement and utilization, suggesting that quick follow-up after agreement is associated with card utilization.

In the not utilized group, the absence of large gaps suggests that factors other than delay between agreement and utilization are influencing the decision not to use the card.

Time Analysis:

The not utilized group had a longer average time between application and agreement compared to the utilized group (9.5 days vs. 7.5 days). This may indicate that a longer decision-making period could correlate with non-utilization.



Utilized Data:

Large Gap (More than 7 Days) Proportion:

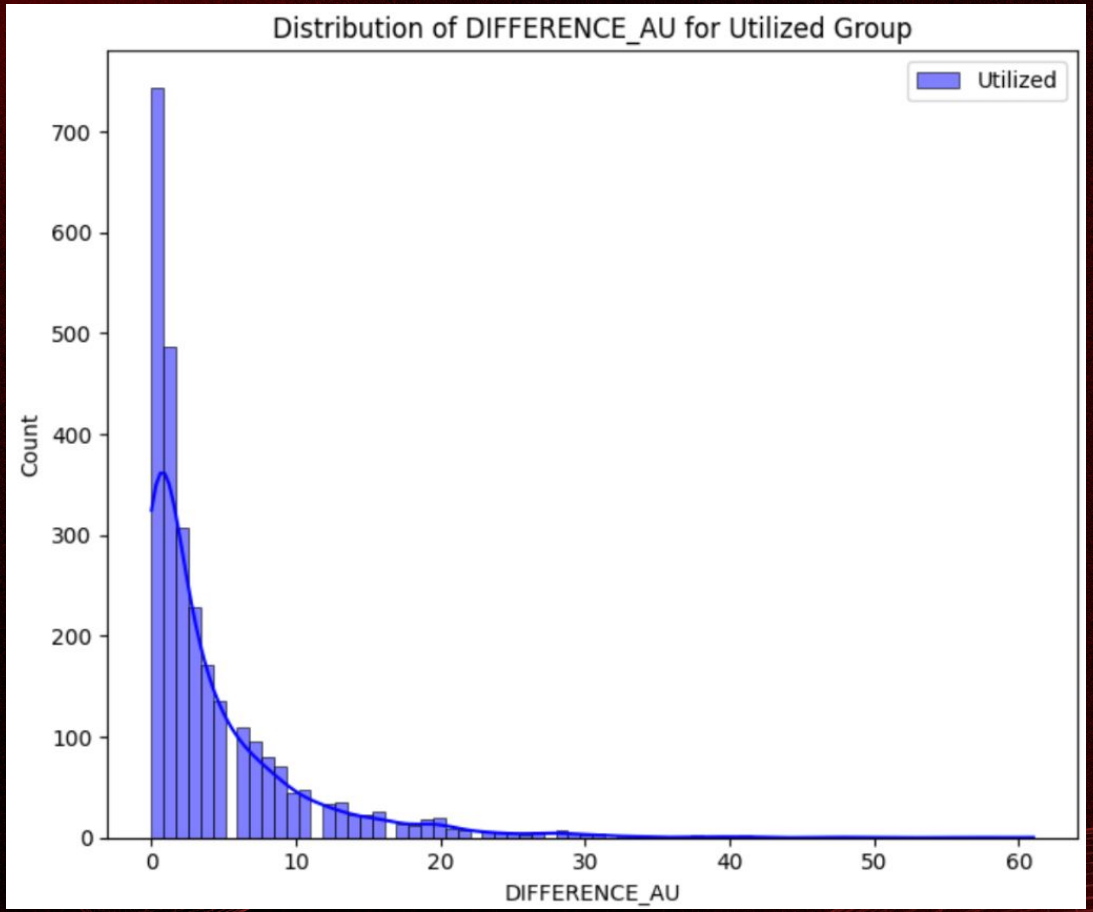
81.42% of the utilized cases had no significant delay between agreement and utilization.

18.58% of the utilized cases experienced a large gap between agreement and utilization.

Not Utilized Data:

Large Gap Proportion:

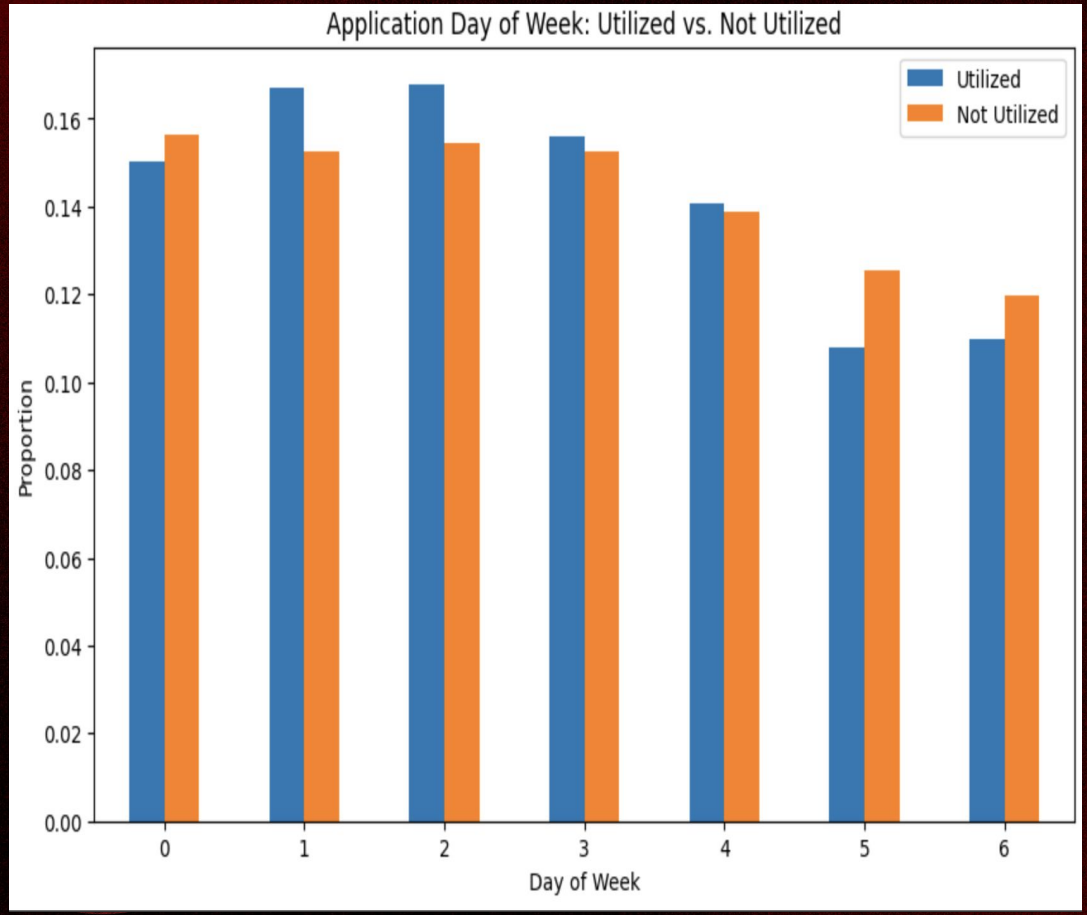
100% of the cases in the not utilized group had no significant gap between agreement and utilization.



Initial Analysis

Balanced Usage

For most days of the week, the proportions of utilized versus not utilized are quite close, indicating that usage is relatively consistent across the week.



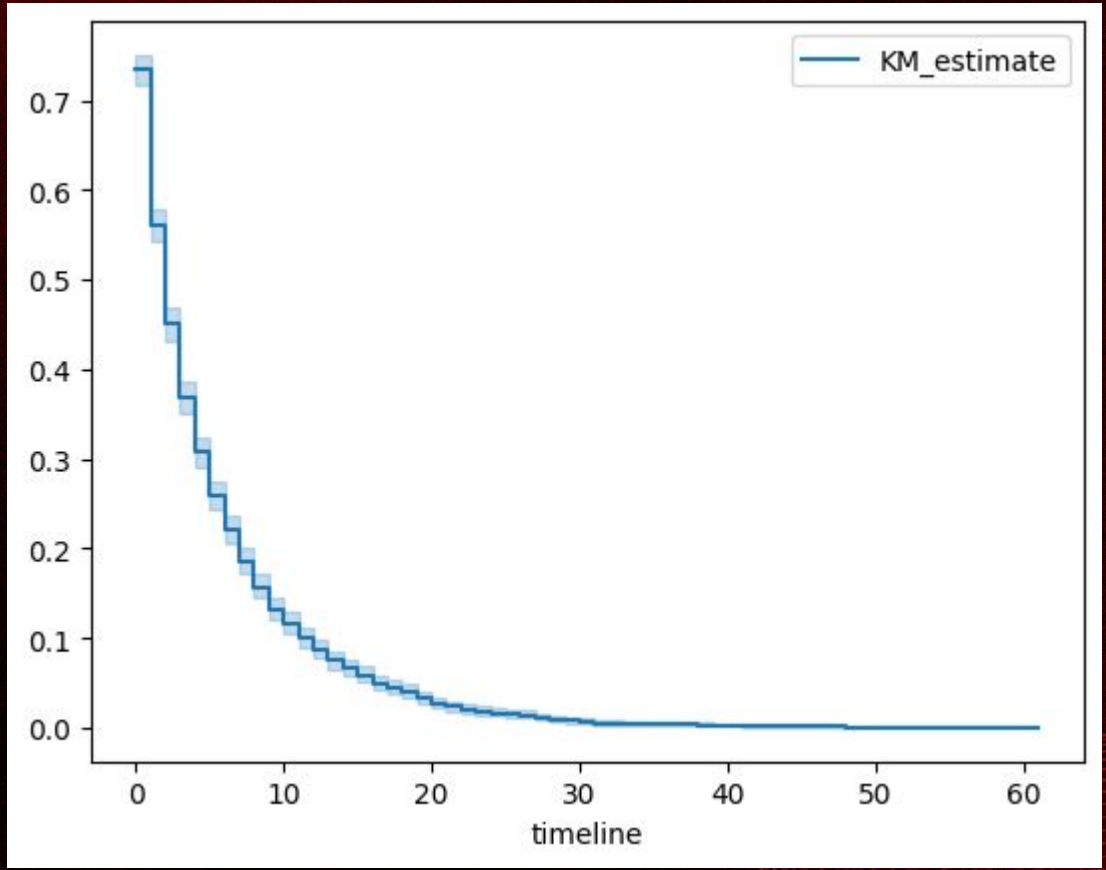
Survival Analysis

Kaplan-Meier curve

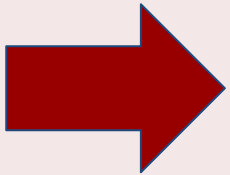
The Kaplan-Meier curve is a visual representation of the survival function, which estimates the probability of an individual surviving (not utilizing the card) past a specific time point

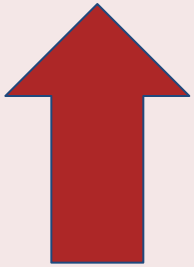
Decreasing curve

Indicates that the probability of not utilizing the card decreases over time, meaning more customers are utilizing their cards.



PREDICTION





To make a forecast until 31/07/2023 for a number of utilizations per day with confidence intervals

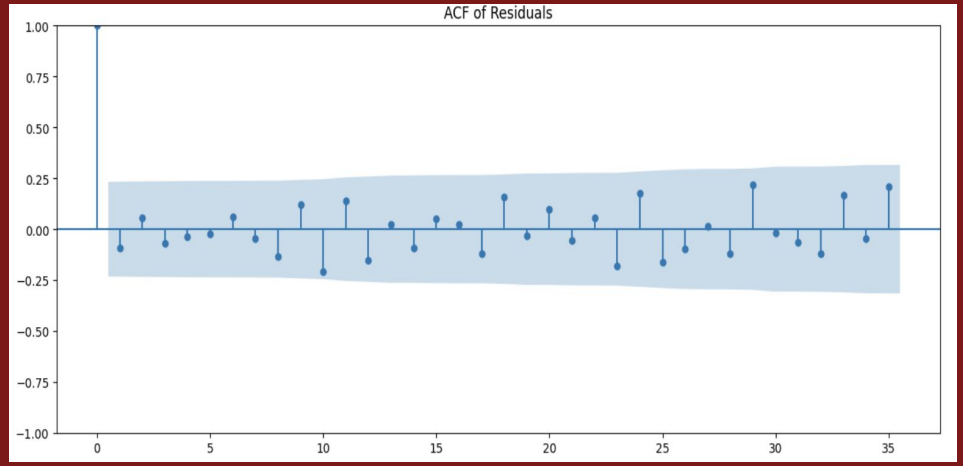
- Time Series Forecasting
- Machine Learning Approach
- Classification with Cohort Analysis Indirect Approach

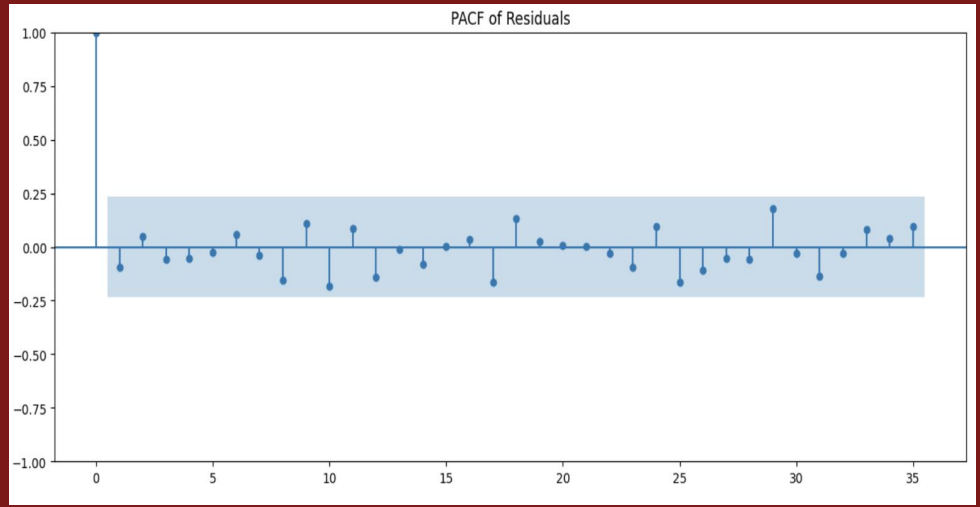
TIME SERIES PREDICTION

The ACF (Autocorrelation Function) plot shows the correlation of the residuals with their own lagged values.

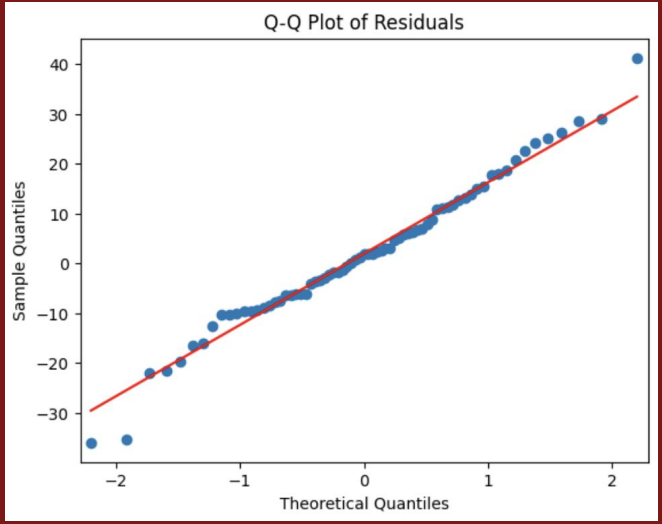
The PACF (Partial Autocorrelation Function) plot shows the partial correlation of the residuals with their own lagged values, controlling for the values of the intermediate lags.

Both ACF and PACF plots indicate that there is no significant autocorrelation remaining in the residuals. This is a good sign, as it suggests that the SARIMA model is well-specified and has effectively captured the patterns in the data.





RESULTS OF TESTS



Ljung-Box test

The Ljung-Box test checks if there is any autocorrelation in the residuals.The p-value of 0.619583 is well above 0.05, indicating that the residuals from your SARIMA model are not significantly autocorrelated, which is a good sign.

The Shapiro-Wilk test

The Shapiro-Wilk test assesses whether the residuals follow a normal distribution. The p-value of 0.6626700162887573 indicates that the residuals are likely normally distributed, which is an assumption in many time series models, including SARIMA.

The Durbin-Watson

The Durbin-Watson statistic checks for the presence of autocorrelation in the residuals. A value of 2.111075109488394 is very close to 2, indicating that there is little to no autocorrelation in the residuals, which is desirable for a well-fitted model.

SARIMAX

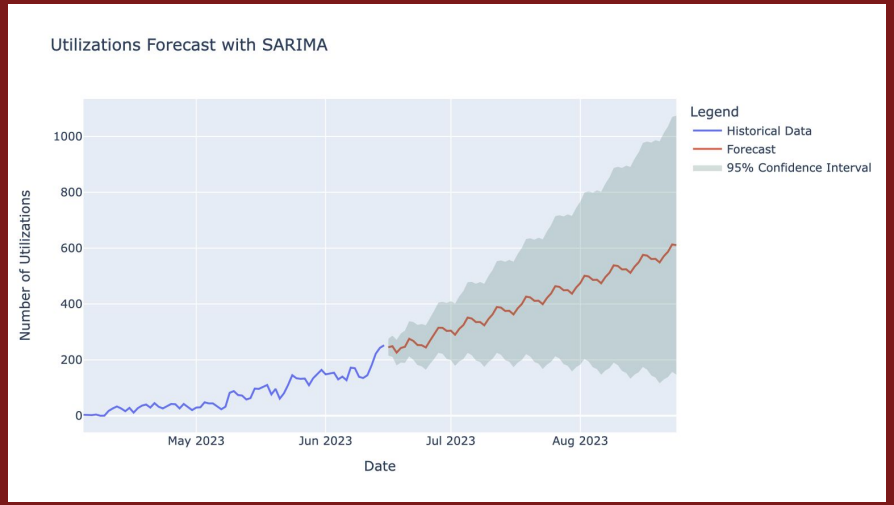
SARIMAX Model for Daily Utilization

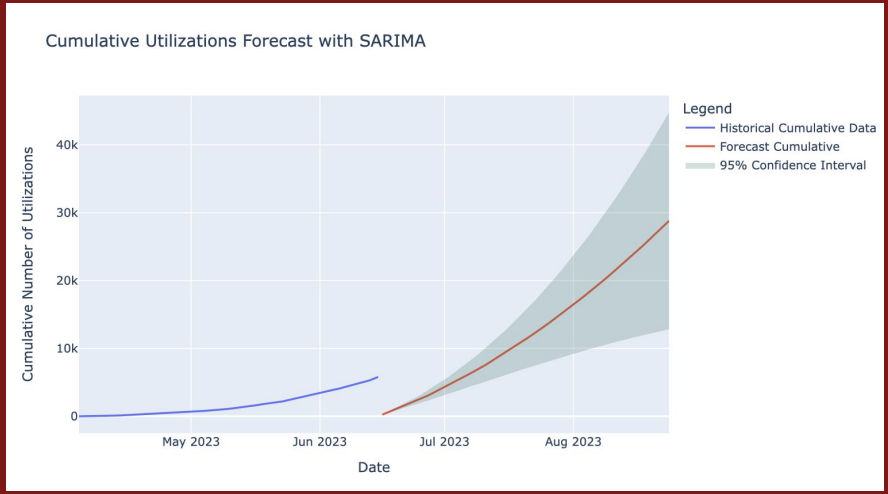
The SARIMAX model was defined and fitted to the daily utilization data.
The `order` parameter was set to `(1, 1, 1)` for non-seasonal components, and `seasonal_order` was set to `(1, 1, 1, 7)` to capture weekly seasonality.

Forecasting

Forecasts were generated for the next 70 days. 95% confidence intervals were computed for the forecasts.

Mean Squared Error (MSE): 200.533
Mean Absolute Error (MAE): 10.957

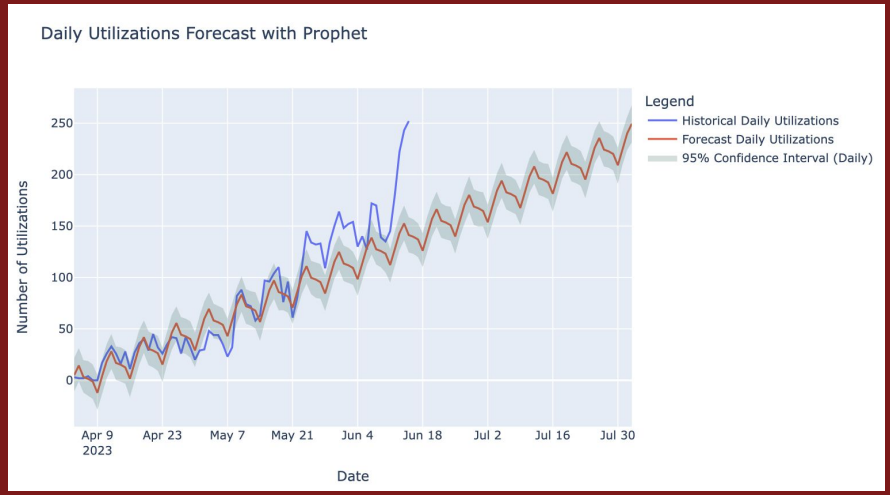


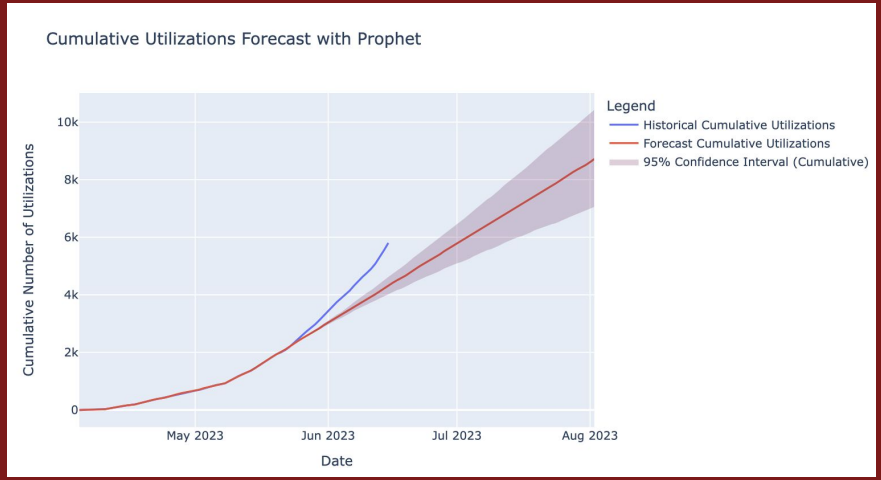


PROPHET

The model's MAE and MSE suggest that the Prophet model performs reasonably well in predicting daily utilizations, though there is some level of error, as reflected in the MAE of 32.

For cumulative utilizations, the errors are larger due to the nature of cumulative data (errors compound over time). The MAE of 385.78 suggests that the model tends to deviate more significantly from the actual cumulative totals, which is normal given the cumulative nature of the data.





Classification with Cohort Analysis

Outline of the Approach

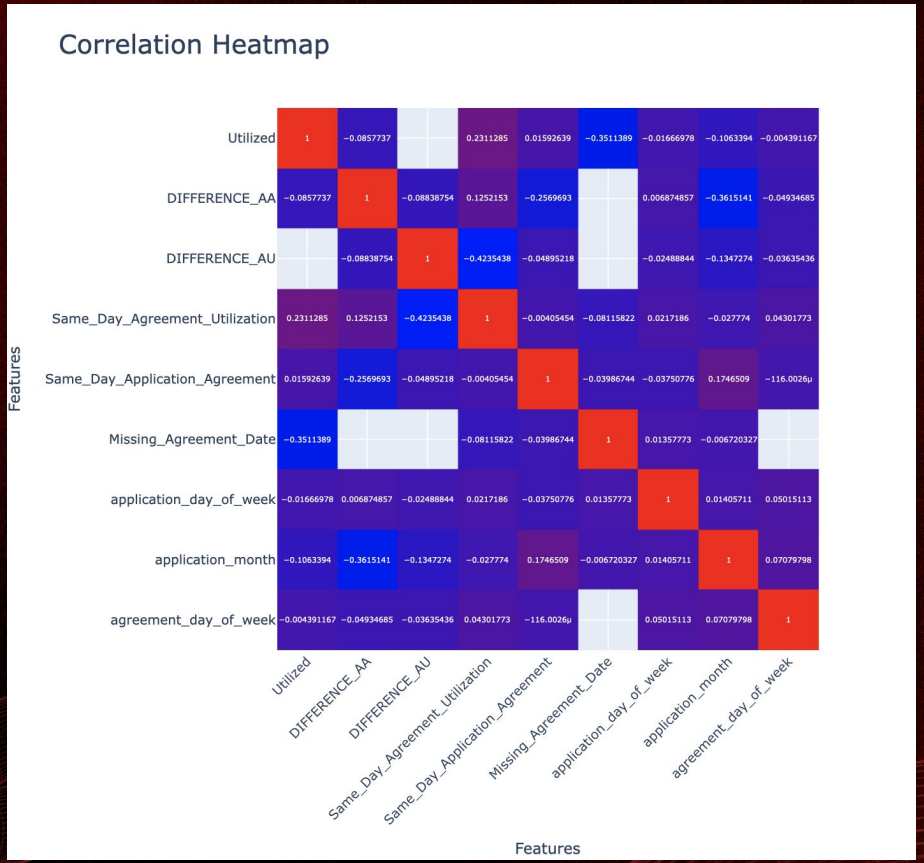
The cohorts based on application and agreement dates were used

Then a classification model was build l to predict if a cohort will utilize the card

The probabilities were converted to expected utilizations.

Expected utilizations were aggregated to obtain daily forecasts

If_utilised column was created to indicate whether the card was used or not



Initial Analysis

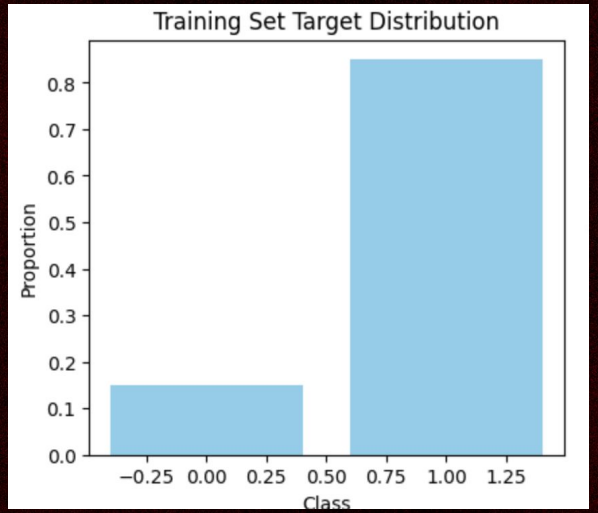
Target variable creation:

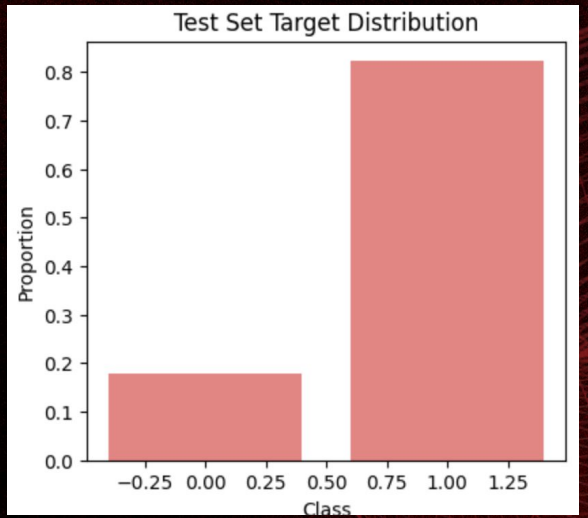
Defining 'Utilized' as a binary indicator of whether a card was utilized.

Feature engineering:

Calculating date differences, creating binary indicators for specific date relationships, and extracting categorical features from dates.

From the picture it can also be seen that the data is imbalance that we will take further into account





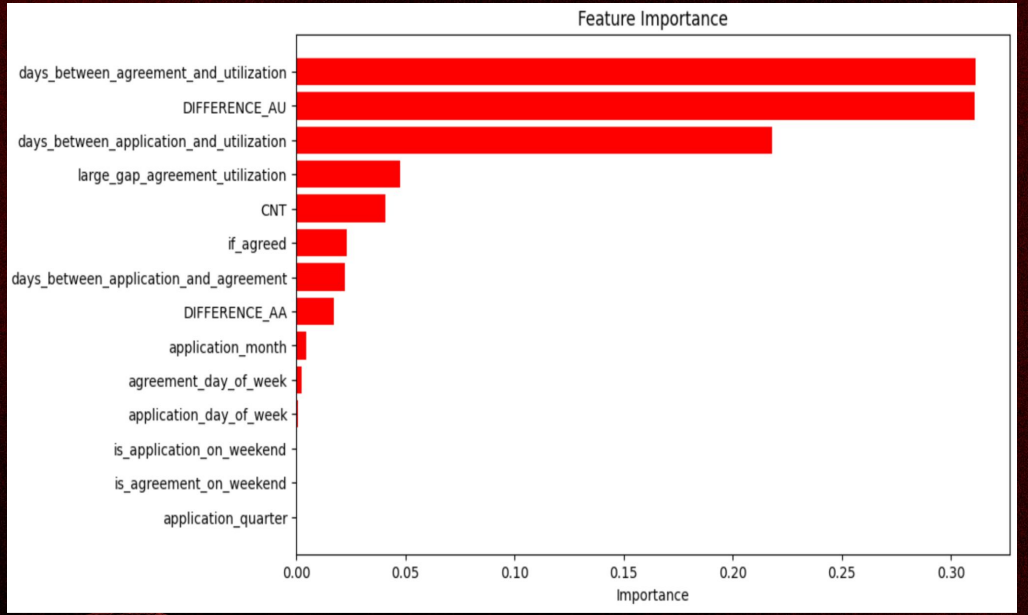
Feature Importance

In feature importance analysis, we focused on evaluating the impact of various features on the performance of a Random Forest model for predicting credit card utilization

Utilized: This feature has the highest importance score, indicating that whether the card was utilized or not is a significant predictor of future utilization.

DIFFERENCE_AU: The difference between application and utilization dates plays a substantial role, with the second-highest importance score.

Other Features: Features like Missing_Agreement_Date, Same_Day_Agreement_Utilization, and day-of-week indicators have lower importance scores, suggesting they contribute less to the model's predictions.

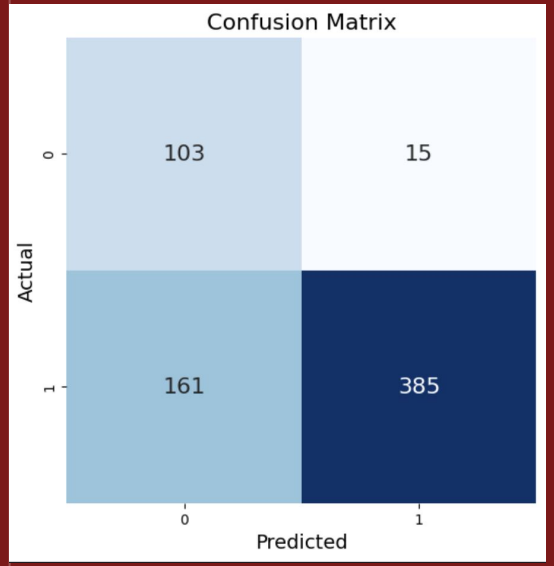


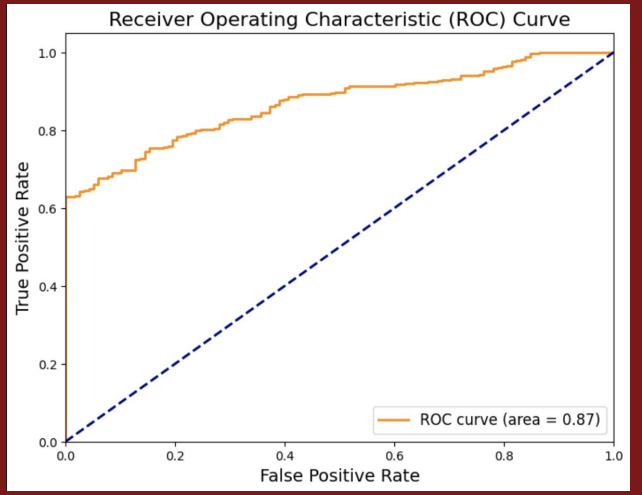
Logistic Regression Baseline

Class 0 (Negative Class): The model has a high recall (0.87) but low precision (0.39), indicating it correctly identifies most of the negatives but with many false positives.

Class 1 (Positive Class): The model has a high precision (0.96) and a moderate recall (0.71), showing it accurately identifies positives with fewer false positives.

To solve the imbalance class problems we defined higher weight to class with less observations



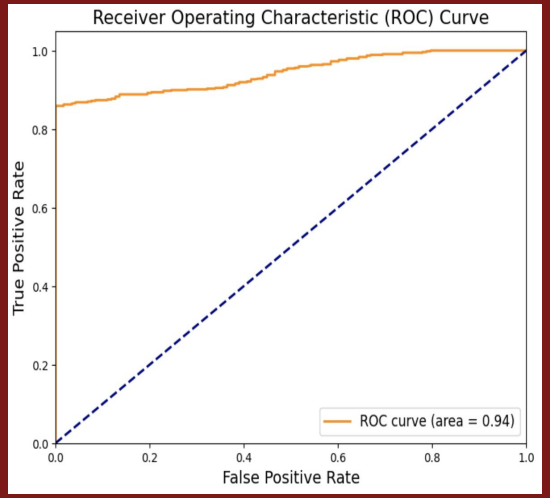


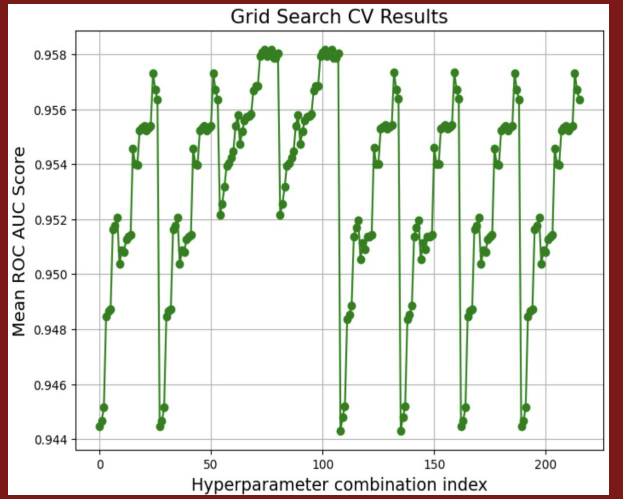
Random Forest Model

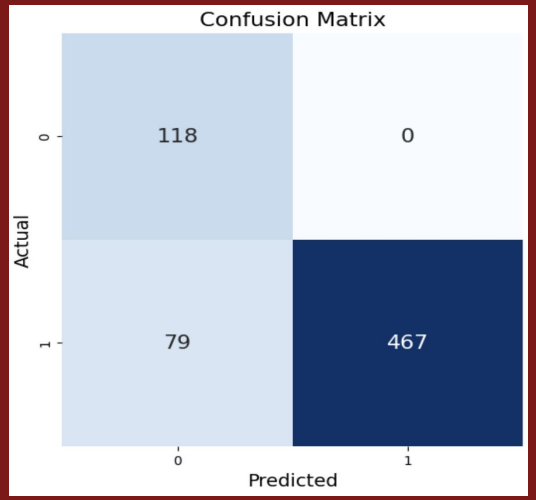
Class 0: The model achieves perfect recall (1.00), meaning it identifies all actual negatives, but its precision is lower (0.60), suggesting it occasionally misclassifies some positives as negatives.

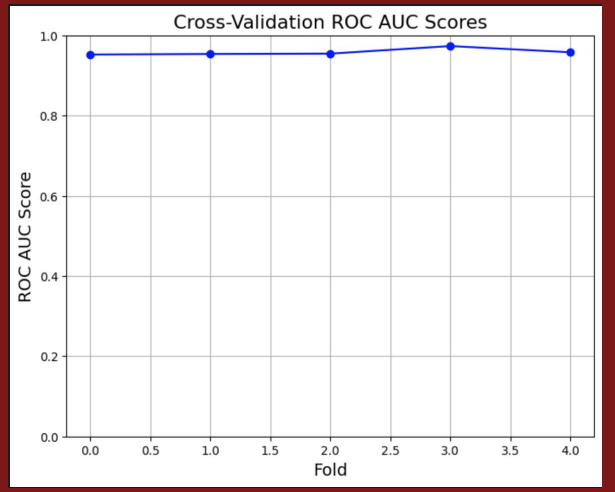
Class 1: The model has high precision (1.00) and strong recall (0.86), indicating it effectively identifies positive cases with fewer false positives.

The model demonstrates high overall accuracy (88%) and performs well in distinguishing between the classes, with better precision for positives and perfect recall for negatives.









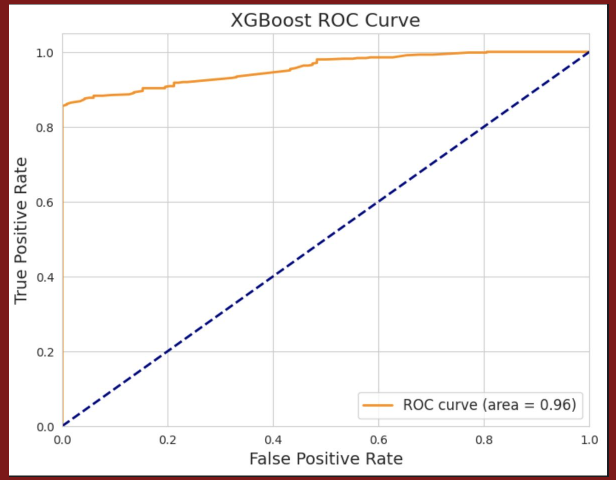
XGBOOST

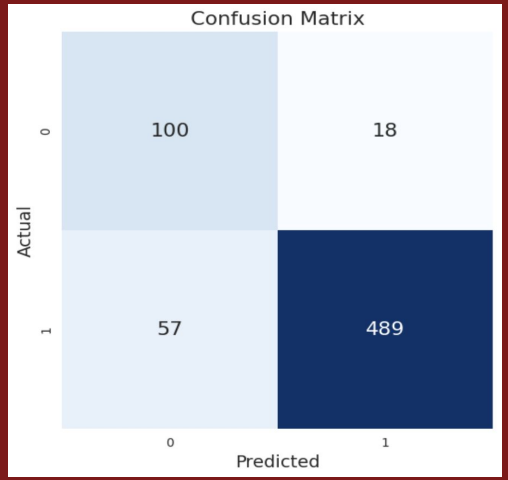
The ROC AUC score measures the model's ability to discriminate between classes. ROC AUC Score: 0.9555

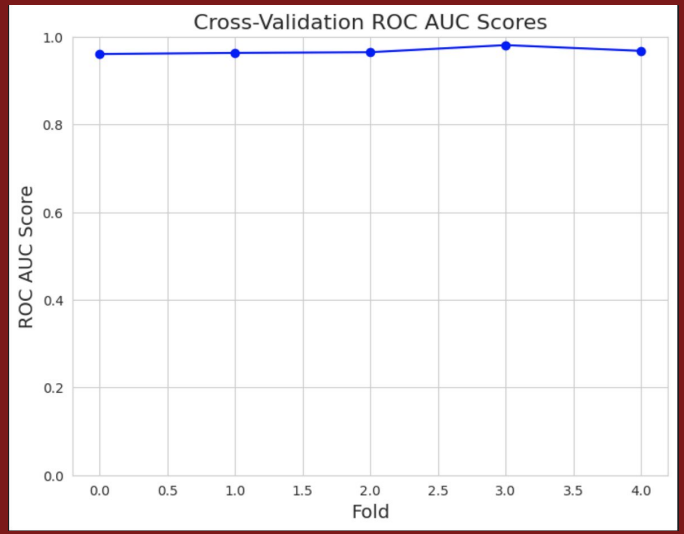
This score indicates an excellent ability to distinguish between the positive and negative classes, with 95.55% of the area under the ROC curve representing the model's performance.

The cross-validation ROC AUC scores provide insight into the model's performance stability]

Mean CV ROC AUC Score: 0.9670



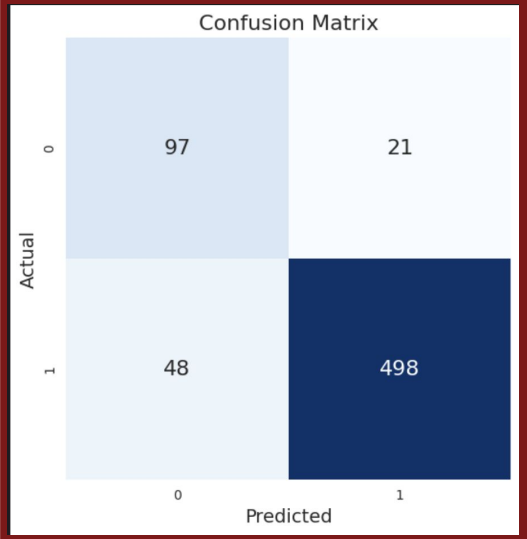




CATBOOST

This score indicates excellent performance, with 95.86% of the area under the ROC curve representing the model's ability to distinguish between positive and negative classes.

The CatBoost model performs exceptionally well with high precision and recall for both classes and exhibits consistent performance across different data splits.



Class 1 (Positive Class):

Precision: 0.96, meaning that 96% of the predicted positives are true positives.

Recall: 0.91, indicating that 91% of actual positives are correctly identified.

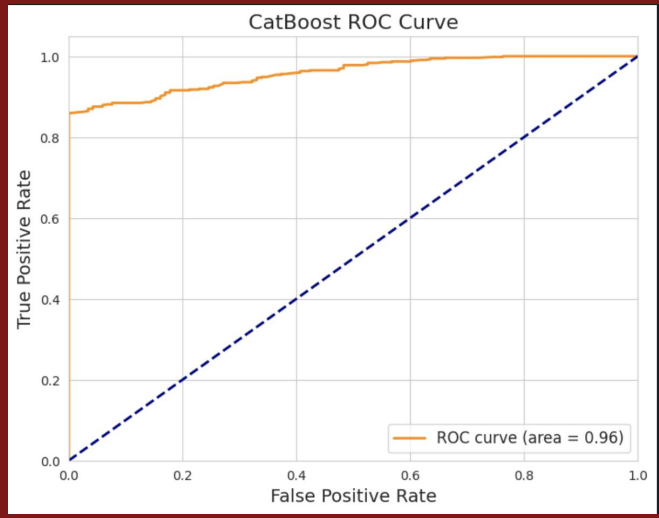
F1-Score: 0.94, good performance in identifying positive cases.

Class 0 (Negative Class):

Precision: 0.67, meaning that 67% of the predicted negatives are true negatives.

Recall: 0.82, indicating that 82% of actual negatives are correctly identified.

F1-Score: 0.74, reflecting a good balance between precision and recall for the negative class.



Combining with Time Series Forecasting

While this approach provides insights into cohort-level utilization, it might not capture daily fluctuations effectively. To receive the results required by the task it is needed to combine with a time series model:

The predicted utilization probabilities can be used as features in the time series model.

Weight the time series forecasts based on the predicted probabilities.

But this approach needs to be developed more so here is the mention of it