Data Mining/Machine Learning Project: Medical Appointments - No Show

Abstract

Missed medical appointments pose a significant challenge to healthcare institutions, leading to inefficient use of resources and potential negative impacts on patient health outcomes. This project aims to develop predictive models to identify patients at risk of missing their appointments. Using a dataset containing medical appointment records, we analyze various factors contributing to no-show behavior and evaluate the effectiveness of logistic regression and random forest classifiers in predicting missed appointments. Our findings indicate that specific factors, such as SMS reminders and the time gap between scheduling and the appointment, significantly influence attendance rates. By incorporating these insights, healthcare providers can improve intervention strategies to reduce missed appointments and enhance service efficiency.

I. Introduction

Missed medical appointments are a pervasive issue in healthcare systems worldwide, resulting in wasted resources and suboptimal patient care. Understanding the underlying factors that contribute to no-shows is critical for developing effective strategies to mitigate this problem. This project leverages data mining and machine learning techniques to predict whether a patient will attend their scheduled appointment. The primary objectives are to identify key factors influencing no-show behavior and compare the performance of logistic regression and random forest classifiers in predicting appointment attendance.

II. Business Understanding

Healthcare institutions incur significant costs due to missed appointments, which can disrupt schedules and reduce the efficiency of medical services. By accurately predicting no-shows, hospitals can take proactive measures, such as sending reminders or rescheduling appointments, to improve attendance rates. Understanding the effectiveness of these interventions, particularly SMS reminders, can further inform appointment management strategies. The goals of this project are to predict no-shows based on various attributes and to identify the most influential factors contributing to missed appointments.

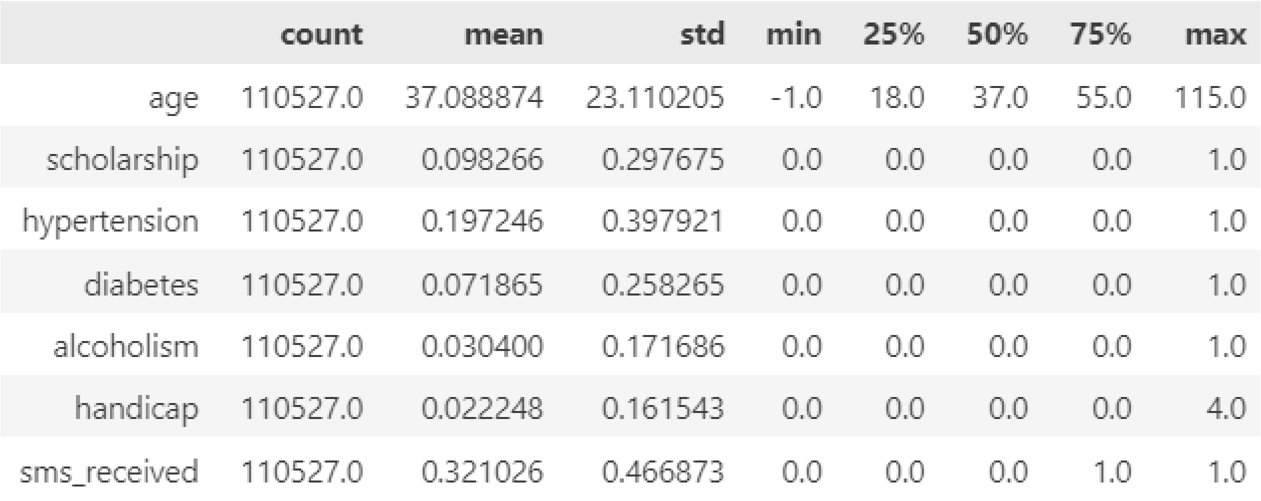
III. Data Understanding

Dataset Description

The dataset comprises 110,527 records of medical appointments with 14 attributes: PatientId, AppointmentID, Gender, ScheduledDay, AppointmentDay, Age, Neighborhood, Scholarship, Hypertension, Diabetes, Alcoholism, Handicap, SMSReceived, and NoShow. The target variable is NoShow, indicating whether a patient attended their appointment.

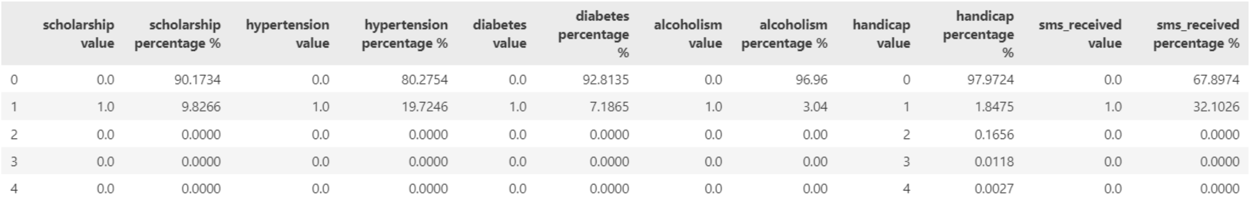
Initial Data Analysis

1. **Data Types and Structure:** The dataset includes both nominal/categorical and discrete/continuous variables. Nominal variables include PatientId, AppointmentID, Gender, Neighborhood, Scholarship, Hypertension, Diabetes, Alcoholism, Handicap, SMSReceived, and NoShow. Continuous variables include Age. ScheduledDay and AppointmentDay are date types.
2. **Missing Values:** There are no missing values in the dataset.
3. **Descriptive Statistics:** Initial analysis revealed anomalies, such as an age value of -1 and inconsistent values in the Handicap column, which required data cleaning.



	count	mean	std	min	25%	50%	75%	max
age	110527.0	37.088874	23.110205	-1.0	18.0	37.0	55.0	115.0
scholarship	110527.0	0.098266	0.297675	0.0	0.0	0.0	0.0	1.0
hypertension	110527.0	0.197246	0.397921	0.0	0.0	0.0	0.0	1.0
diabetes	110527.0	0.071865	0.258265	0.0	0.0	0.0	0.0	1.0
alcoholism	110527.0	0.030400	0.171686	0.0	0.0	0.0	0.0	1.0
handicap	110527.0	0.022248	0.161543	0.0	0.0	0.0	0.0	4.0
sms_received	110527.0	0.321026	0.466873	0.0	0.0	0.0	1.0	1.0

Table 1. High-Level Descriptive Statistics of Features



	scholarship value	scholarship percentage %	hypertension value	hypertension percentage %	diabetes value	diabetes percentage %	alcoholism value	alcoholism percentage %	handicap value	handicap percentage %	sms_received value	sms_received percentage %
0	0.0	90.1734	0.0	80.2754	0.0	92.8135	0.0	96.96	0	97.9724	0.0	67.8974
1	1.0	9.8266	1.0	19.7246	1.0	7.1865	1.0	3.04	1	1.8475	1.0	32.1026
2	0.0	0.0000	0.0	0.0000	0.0	0.0000	0.0	0.00	2	0.1656	0.0	0.0000
3	0.0	0.0000	0.0	0.0000	0.0	0.0000	0.0	0.00	3	0.0118	0.0	0.0000
4	0.0	0.0000	0.0	0.0000	0.0	0.0000	0.0	0.00	4	0.0027	0.0	0.0000

Table 2. Percentage class distribution per categorical feature

IV. Data Wrangling/Preliminary Cleaning

Data cleaning involved the following steps:

1. **Removing Anomalies:** The record with an age of -1 was removed.
2. **Converting Date Types:** The ScheduledDay and AppointmentDay columns were converted to datetime format to facilitate time-related calculations.
3. **Handling Inconsistent Values:** 199 rows containing inconsistent values for handicap were dropped reducing the total records to 110,327.

V. Exploratory Data Analysis

Age Group Distribution

The age distribution revealed a higher frequency of appointments for young people with a left-skewed distribution showing fewer elderly patients (75-100).

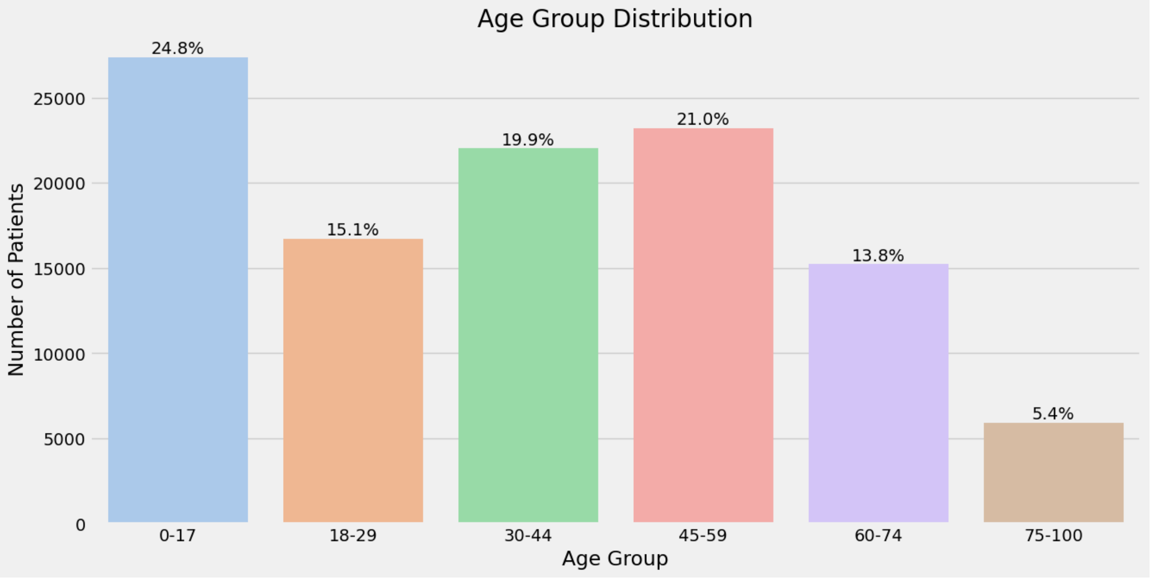


Figure 1. Age group distribution

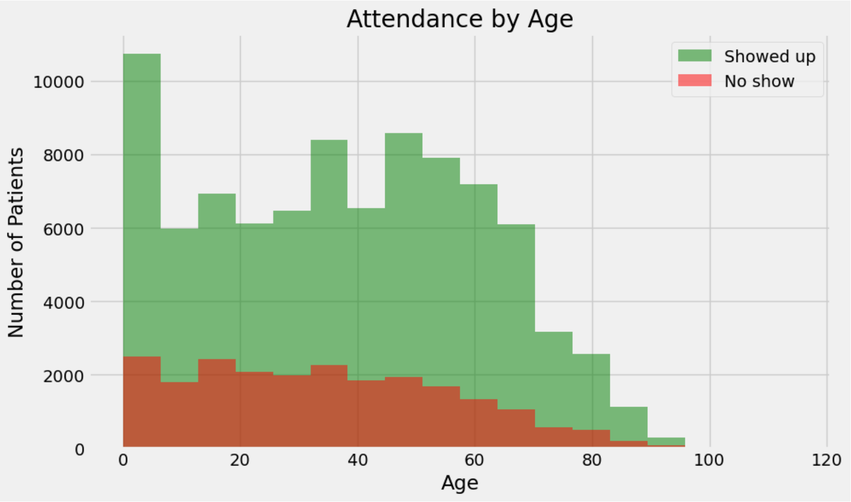


Figure 2. Attendance based on age.

Gender Comparison

Females constituted much of the dataset, with attendance and no-show rates similar across genders, suggesting that gender is not a strong predictor of no-show behavior.

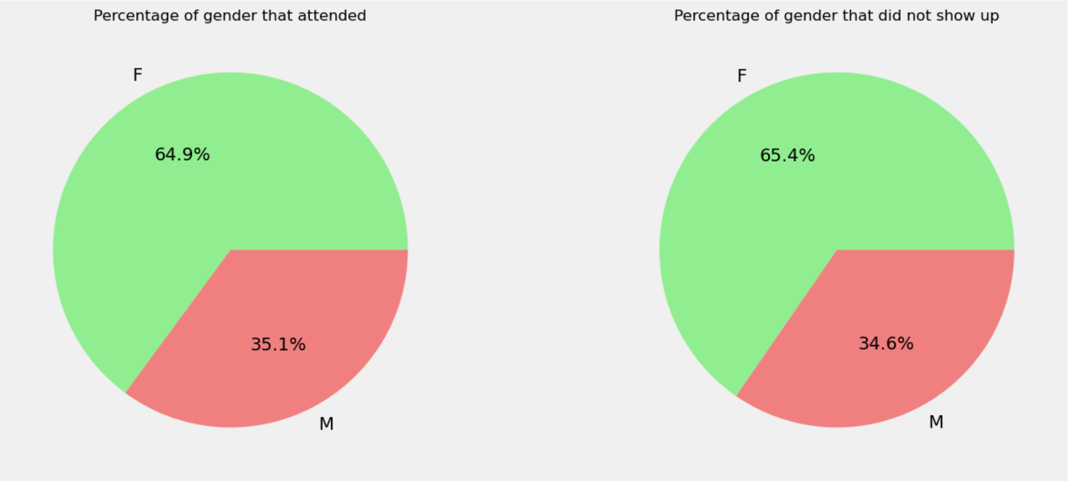


Figure 3. Attendance based on gender.

Chronic Diseases

In exploring the correlation between chronic diseases and appointment attendance, our objective is to understand whether patients with chronic conditions may demonstrate distinct attendance patterns compared to those without such ailments. The analysis unveils a noticeable contrast in attendance rates, with 82.23% of patients with chronic diseases attending appointments versus 79.09% of those without. Conversely, 17.77% of patients with chronic diseases missed appointments, while 20.91% of those without chronic diseases did.

This disparity, a 3.14% difference in attendance rates, although may be thought of being relatively small, slightly suggests that ongoing health management may influence attendance behaviour, providing insights for healthcare providers to tailor interventions and enhance appointment adherence across patient groups. However, the scale of this influence may not be determined yet as the records occurred in a short time (40 days). A longer time frame collection may yield better clarity in understanding this influence. But for the goal of the data exploration and modelling, chronic diseases such as hypertension, diabetes and alcoholism in this dataset do not show a noteworthy relationship with appointment adherence.

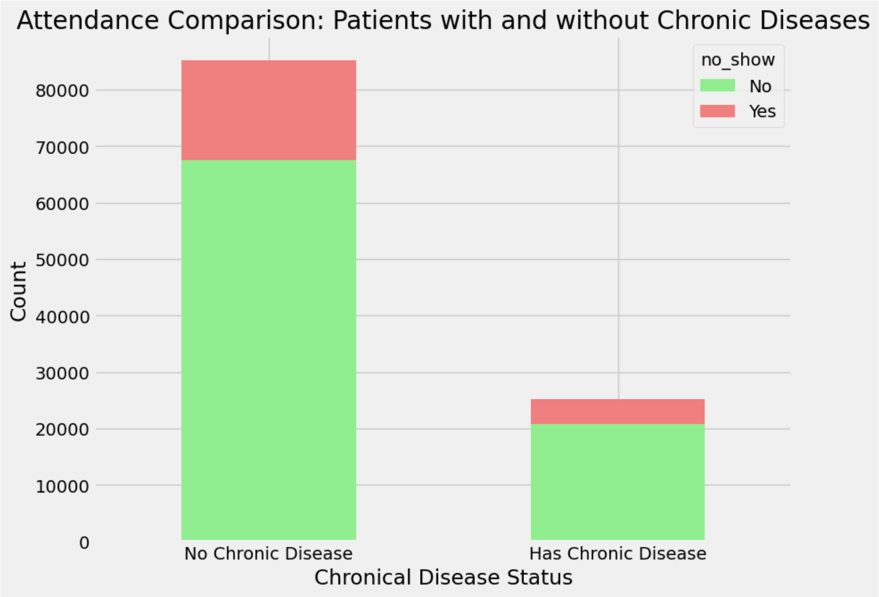


Figure 4. Attendance comparison based on patients' chronic disease status.

SMS Reminders

In analysing the correlation between SMS reception and appointment attendance, our aim is to discern whether patients who receive SMS reminders exhibit different attendance behaviour compared to those who don't. The results revealed a notable difference in attendance rates: 83.30% of patients who did not receive an SMS reminder attended their appointments, while 16.70% did not. The discrepancy showed that sending SMS reminders had an opposite outcome of the expectation that the reminders would improve attendance. However, it is important to investigate how same-day appointments contributes to these findings.

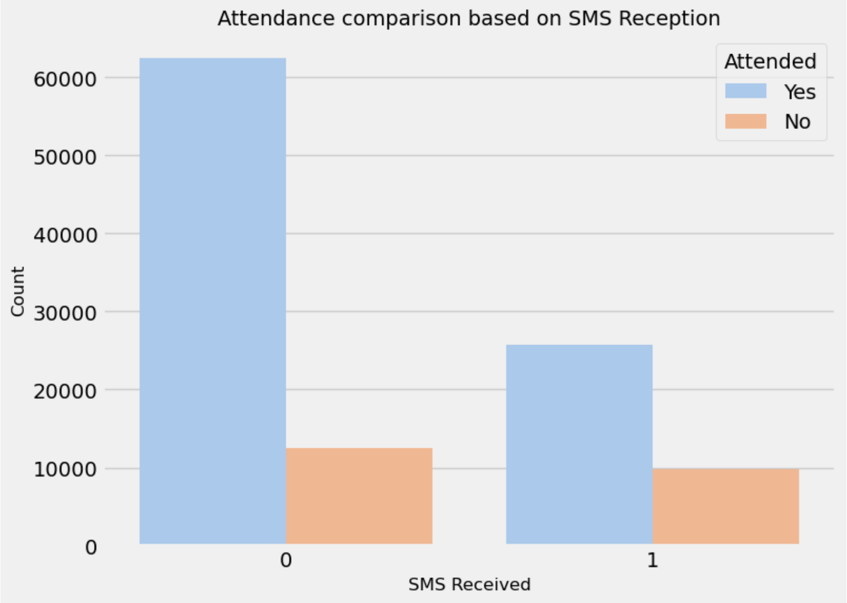


Figure 5. Attendance based on SMS reception.

Roughly 35% of all appointments recorded were same-day appointments. This distribution was significant enough to influence the results gathered earlier. Therefore, it is necessary to filter out same-day appointments as this will be the real test of the impact of the SMS campaign.

After filtering out same-day appointments, the new analysis revealed that patients who did not receive an SMS had a show percentage of 70.55% and a no-show percentage of 29.45%. Those who received an SMS showed a slight increase in attendance, with a show percentage of 72.43% and a no-show percentage of 27.57%. This suggests that, for non-same-day appointments, receiving an SMS has a modest but positive impact on attendance, improving the show rate by approximately 2% compared to those who did not receive an SMS.

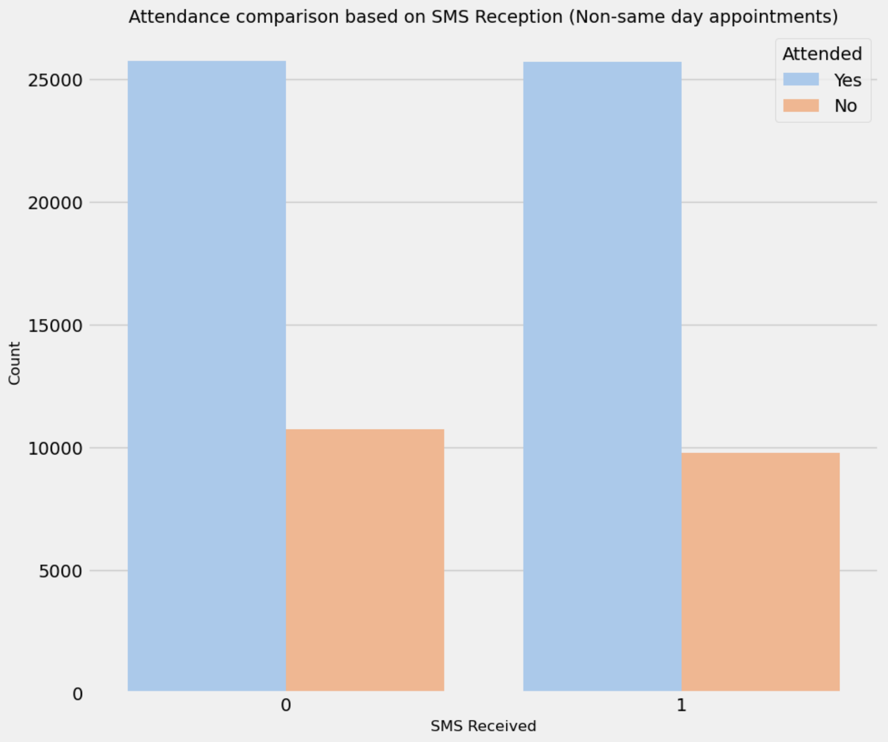


Figure 6. Attendance comparison based on SMS reception (non-same-day appointments).

Handicap Levels

Patients with handicaps had slightly lower no-show rates compared to those with no handicaps, but this inference warrant cautious interpretation as there is a significant imbalance in the distribution, with most patients having no handicaps.

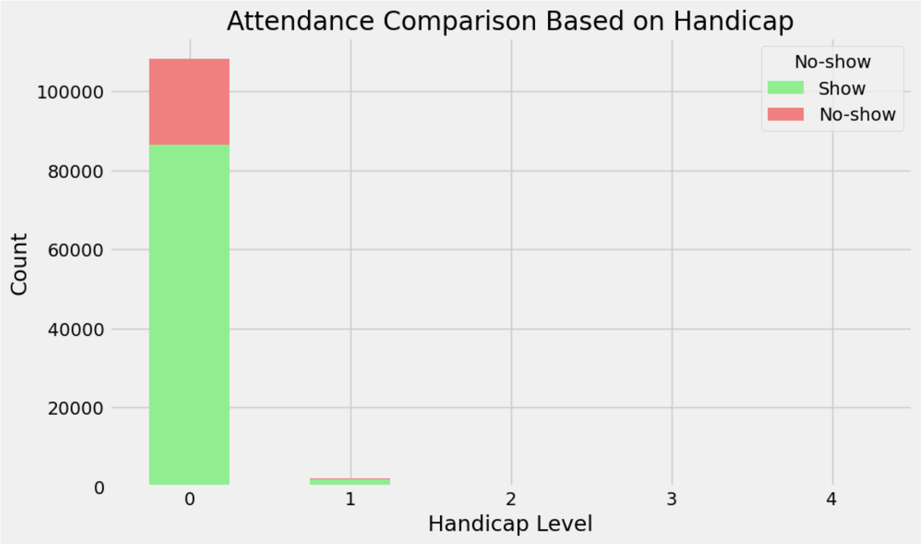


Figure 7. Attendance based on handicap.

Scholarship Status

Patients without scholarship status had a higher attendance rate compared to those with scholarship status suggesting a potential socioeconomic influence on appointment adherence.

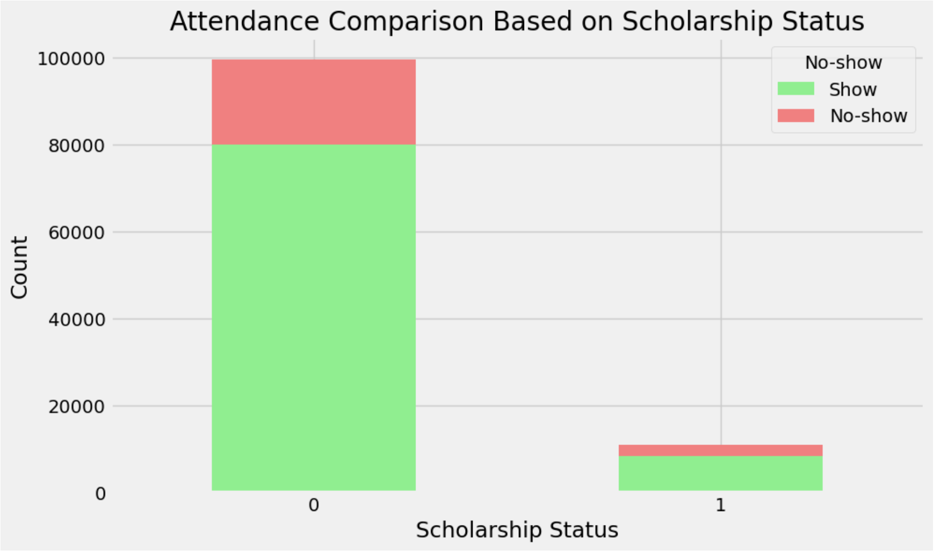


Figure 8. Attendance based on scholarship status.

Neighborhood

Neighborhood had a notable impact on attendance rates, with variability in attendance percentages across different neighborhoods, indicating it as a potentially strong predictor of no-show behavior.

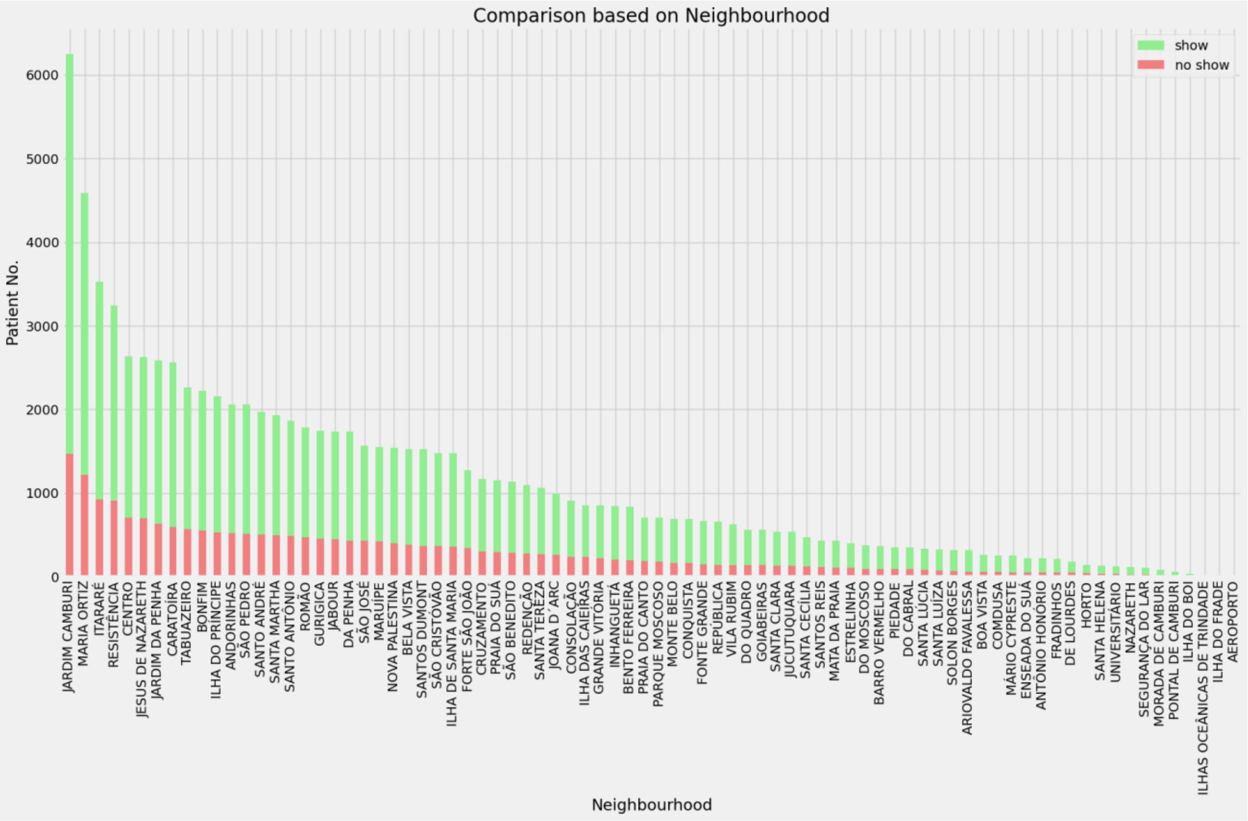


Figure 9. Attendance based on neighborhood.

Class Imbalance

There is a significant imbalance between the classes as over 88k patients attended their appointments versus over 22k missing their appointments. A similar imbalance still appears even after filtering out same-day appointments as it was already known that 35% of the appointments were same-day appointments which majorly were shows (No in no-show class). This occurrence must be considered during data modelling. This also means the metric for evaluating model quality and performance may not be accuracy and might be other metrics like F1 Score and ROC AUC. Another possible technique that can be implemented could be resampling techniques like Random Under Sampling, Random Over Sampling or a hybrid of both (with SMOTE).

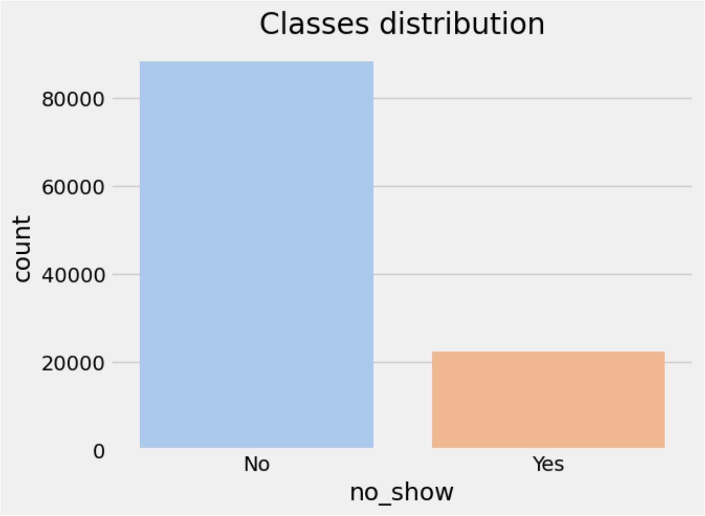


Figure 10. Target variable (No Show) distribution.

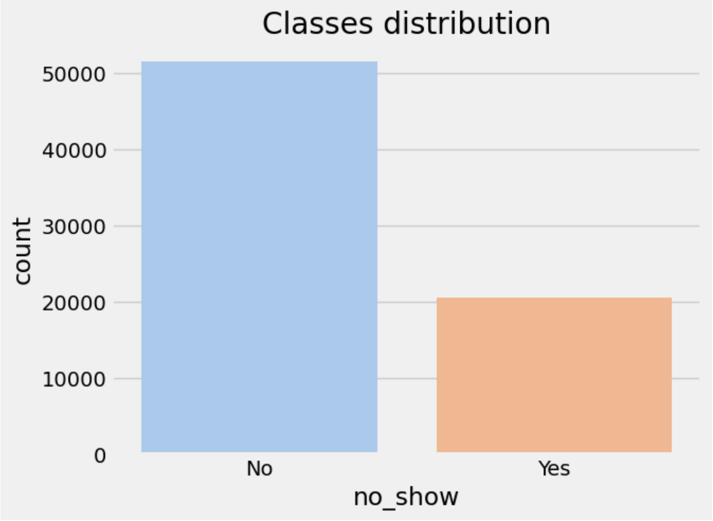


Figure 11. Target variable (No Show) distribution with same day appointments removed.

VI. Data Modelling

The ‘business’ goal for this project is to produce a model that can precisely predict the positive class of the target variable No-Show while maintaining its generalization property. For this project, two models were explored, namely Logistic Regression and Random Forest Classifier.

Logistic Regression

Logistic Regression is a statistical technique primarily used for binary classification tasks, where the outcome variable can take only two possible values. It works by estimating the probability that a given input belongs to one of the two categories. At the core of Logistic Regression is the logistic function, also known as the sigmoid function, which transforms the linear combination of input features into a probability score ranging between 0 and 1. Many data scientists tend to use logistic regression first on a new dataset as it is simple, yet effective and can serve as a baseline for comparison with other more complex models.

$$P(y = 1|x) = 1(1 + e^{-z})$$

Random Forest Classifier

The Random Forest Classifier is a powerful algorithm that creates multiple decision trees and combines their predictions to make accurate classifications. It was selected for this project because it handles both categorical and numerical data well, and its ensemble approach helps prevent overfitting. By analyzing feature importance, it provides insights into the factors influencing appointment attendance, making it a potentially valuable tool for predicting medical appointment no-shows.

Methodology

To effectively build the best classification models, the methods for training, validating, and testing were as follows:

Convert categorical variables to numerical using one-hot encoding.

This step is crucial as the models can only train and predict numerical values. In this case, the categorical variables were binary and easily represented in 0 and 1s. The gender column with M and F strings were converted to binary numbers. However, the neighborhood feature is one with high cardinality. This nature will severely affect the performance of the model and increase the complexity in the data understanding or evaluation of metrics. As a result, frequency encoding, a common technique was implemented on the neighborhood column.

Frequency encoding replaces the column value for all rows with the frequency of the unique categories in the column. This solves the issue with one-hot encoding whereby there would be significantly high dimensionality or the ‘curse of dimensionality’ as it is commonly said.

Feature Scaling

This step constrains the values within a column between 0 and 1. This brings about fairness to the weights of each column during the model’s learning process. The age column stands out especially with values ranging from 0 to 115. To properly scale this column, standardization or normalization are the most common techniques. Min-max scaling was implemented for this phase.

Feature Selection

This is one of the most crucial steps before model training. Before this learning process began, only the features that were evaluated to correlate the most with the target variables were selected to train the model. This improves the effectiveness of the training as this is an attempt to reduce redundancies. In the real world as well, with hundreds or thousands of features, the high dimensionality problem is prevalent. For this project, two techniques were applied to select the features for model training: Chi Square Test, and Pearson Correlation Efficient.

The Chi-square test evaluates how categorical variables are associated, checking if the observed distribution differs significantly from what we'd expect by chance. The higher the Chi-value, the more dependent the target variable is on this independent variable and vice versa for P-value. On the other hand, the Pearson correlation coefficient gauges the strength and direction of the linear relationship between two continuous variables. The value may range from -1 to 1 where both extremes are the extremity of negative or positive correlation.

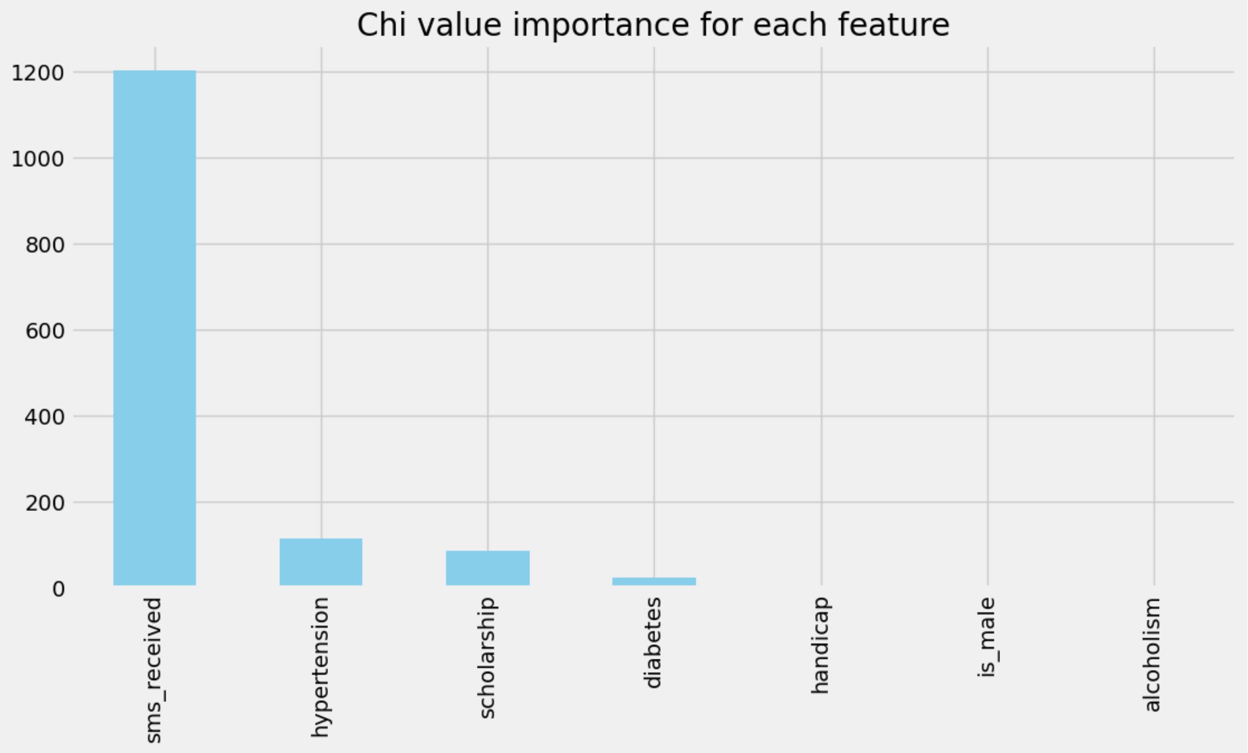


Fig. 12. Chi value importance for categorical features

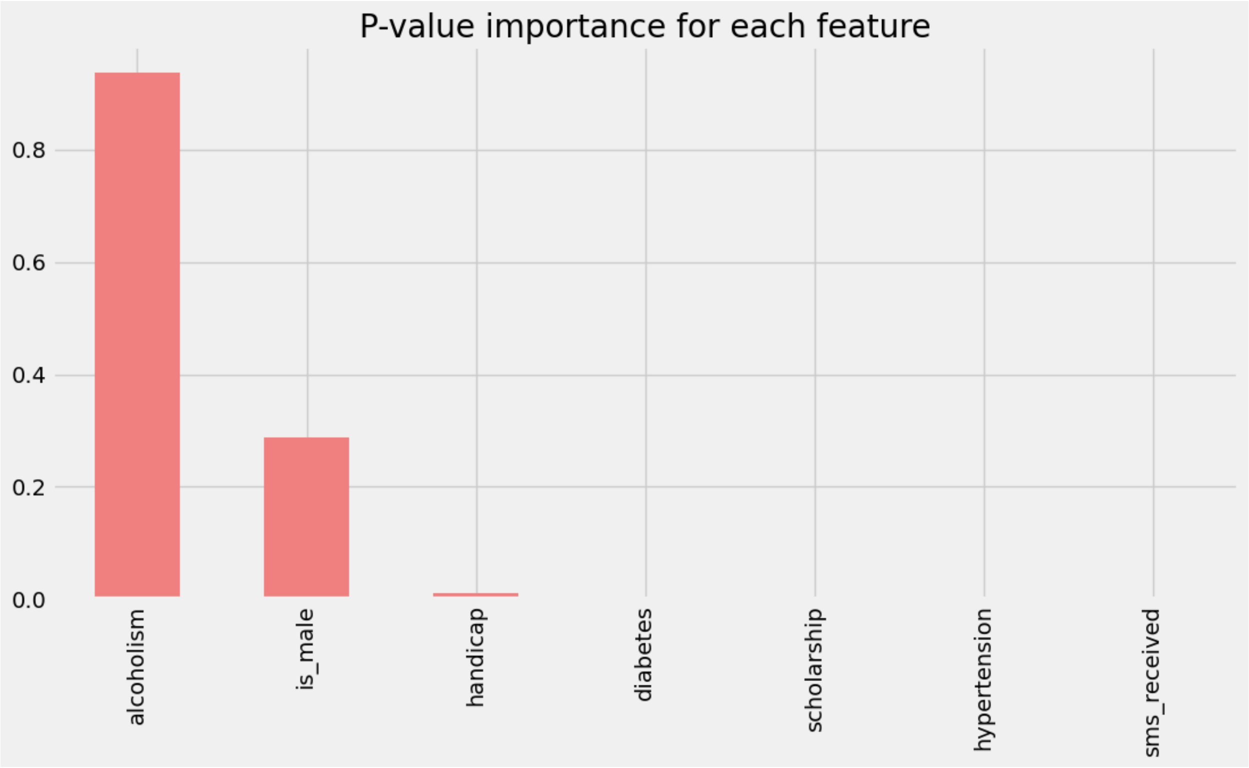


Fig. 13. P-Value importance for categorical features.

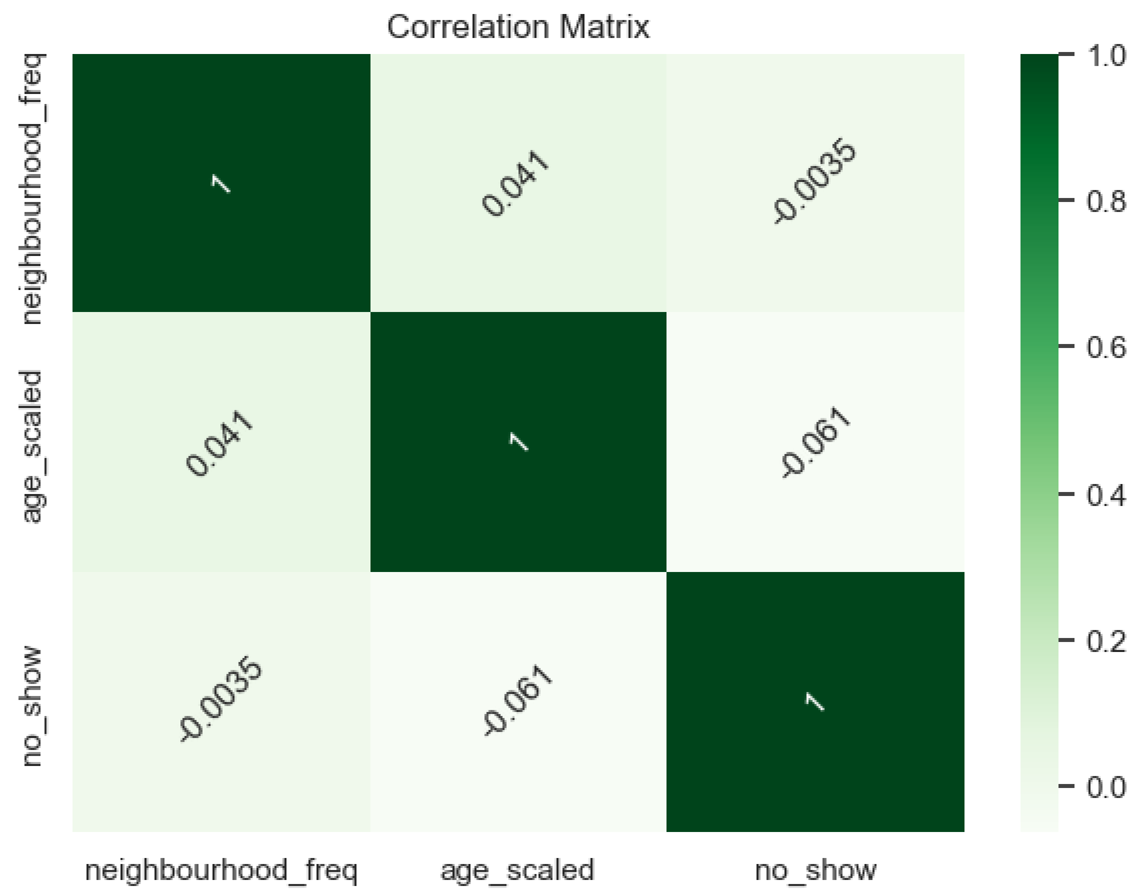


Fig. 14. Pearson Correlation Coefficient heatmap for age and neighborhood.

From conducting these two techniques, three features from the categorical were selected: ‘sms_received’, ‘hypertension’, ‘scholarship’ and ‘neighborhood_freq’ which is the frequency encoded version of the neighborhood feature. The scaled age was also selected.

Metric Selection for determining performance.

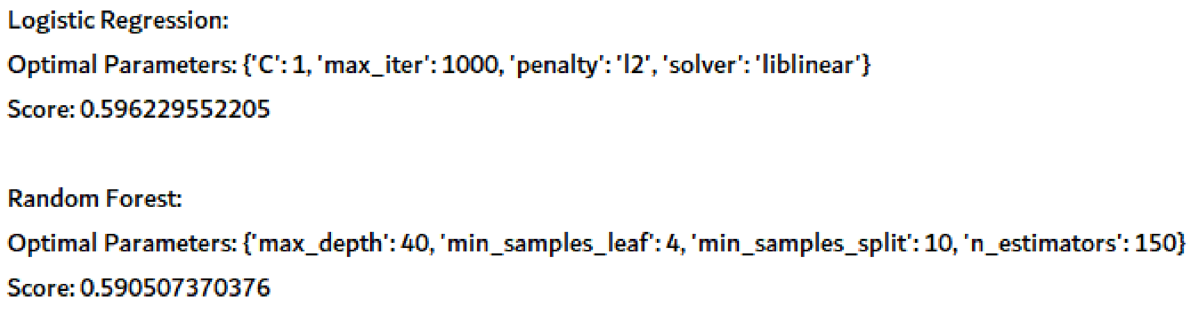
As the accuracy of the model is not to be solely focused on, other metrics need to be considered. For this classification problem where the focus is on the minority class (No show is Yes), the ROC AUC score which represents the area under the ROC curve, explaining the model's ability to distinguish between positive and negative instances across various classification thresholds would be the ideal metric. The ROC AUC score ranges from 0 to 1 with a score of 0.5 meaning random guessing, while 1 indicates perfect prediction. This metric provides a comprehensive outlook of the model's discriminatory power and effectiveness. The F1-Score of the minority class is also important as it is the harmonic mean of the precision and recall.

Stratified K-Fold Cross Validation on Training Set with Logistic Regression and Random Forest

This was implemented via with a stratified K-Fold of 5 number of splits with shuffling. This enables validating of the models on the training consistently as a simple train, validation, test split may generate inconsistent metrics depending on the split distribution per turn.

Hyperparameter Tuning with GridSearchCV

The base logistic regression and random forest models performed poorly especially on metrics for the positive class (No Shows 1). Base Random Forest Classifier performed the worst out of the two, in predicting patients not showing up, which is the main concern for the project. However, we can use GridSearchCV, a method provided by the scikit-learn library that exhaustively runs the base models using all possible combinations of a parameter grid to perform validation. The best model is returned with the optimal hyperparameters.
To perform hyperparameter tuning:
1. Define the hyperparameters to tune for each model
2. Use the training set for
3. Run GridSearchCV cross-validation for each model on the training set with a K-fold of 5 and store the optimal parameters.
4. Store the optimal parameters and their respective scores. These will be determined using multi-metric scoring based on F1-Score, ROC AUC, and accuracy.



Logistic Regression:

Optimal Parameters: {'C': 1, 'max_iter': 1000, 'penalty': 'l2', 'solver': 'liblinear'}

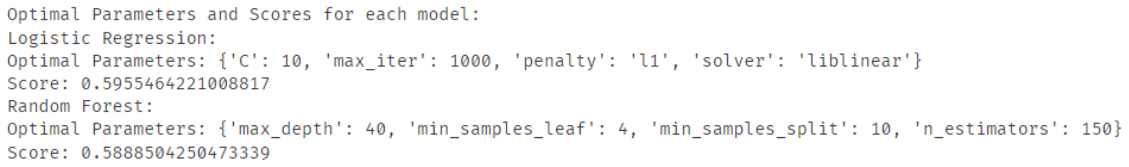
Score: 0.596229552205

Random Forest:

Optimal Parameters: {'max_depth': 40, 'min_samples_leaf': 4, 'min_samples_split': 10, 'n_estimators': 150}

Score: 0.590507370376

With GridSearchCV, it was observed that the best parameters found for the models still yield a poor ROC_AUC score. Therefore, even with hyperparameter tuning, the models do not generalize or predict the positive class well.



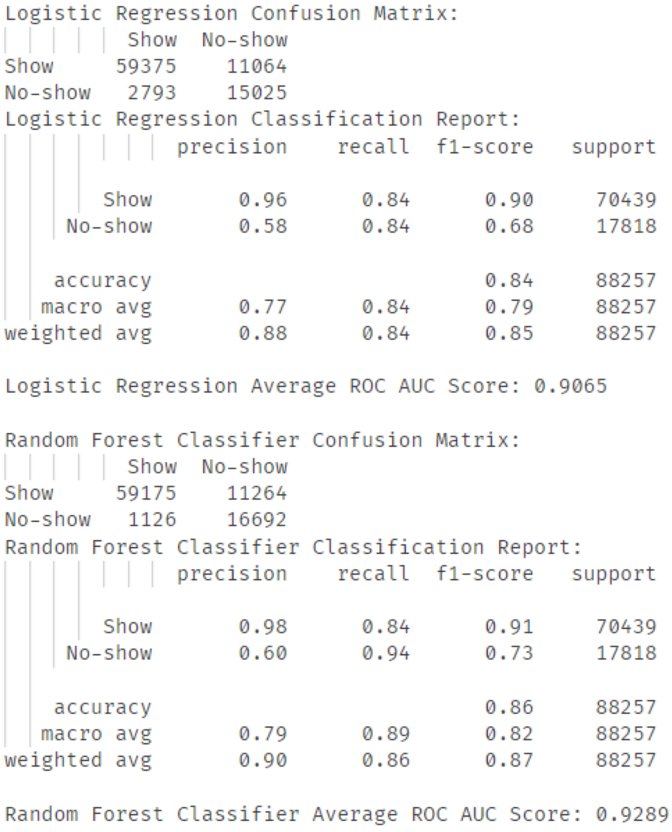
```
Optimal Parameters and Scores for each model:
Logistic Regression:
Optimal Parameters: {'C': 10, 'max_iter': 1000, 'penalty': 'l1', 'solver': 'liblinear'}
Score: 0.5955464221008817
Random Forest:
Optimal Parameters: {'max_depth': 40, 'min_samples_leaf': 4, 'min_samples_split': 10, 'n_estimators': 150}
Score: 0.5888504250473339
```

Feature Engineering
Feature engineering is another way of improving the performance of machine learning models, particularly when initial attempts with other methods have not yielded significant improvements. Below are some proposed new features that could potentially enhance the model's ability to predict no-shows:

1. Days Between Scheduling and Appointment: The time gap between when an appointment is scheduled and when it is held could influence the likelihood of a no-show. Longer gaps may lead to more no-shows due to changes in patients' schedules or forgotten appointments.
2. Previous No-Shows: Patients with a history of no-shows are more likely to miss future appointments. This feature can capture the no-show behavior of patients.

Predict target variable given new features.

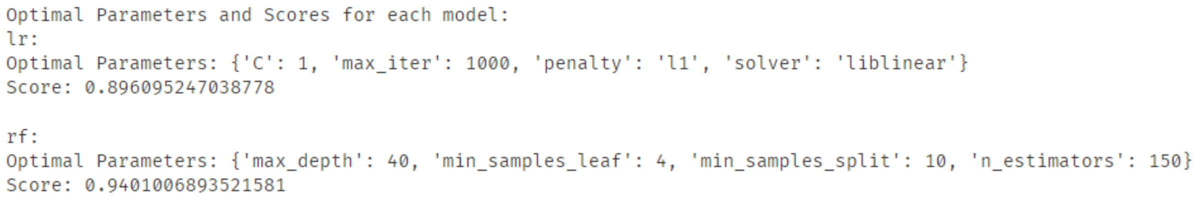
This stage involved predicting the classes using the new features added to the data frame. From the results, a significant improvement was observed after a 5-fold cross validation.
Both models were able to predict No-shows significantly better. This shows a clear positive impact of the derived features via feature engineering. One can attempt to improve this performance even further as although the F1-Score for the No show improved for both models, through a round of hyper parameter tuning, the models might improve even more in these metrics.



```
Logistic Regression Confusion Matrix:
          Show  No-show
Show     59375    11064
No-show   2793    15025
Logistic Regression Classification Report:
              precision    recall  f1-score   support

        Show       0.96      0.84      0.90     70439
     No-show       0.58      0.84      0.68     17818

    accuracy                           0.84     88257
   macro avg       0.77      0.84      0.79     88257
weighted avg       0.88      0.84      0.85     88257

Logistic Regression Average ROC AUC Score: 0.9065

Random Forest Classifier Confusion Matrix:
          Show  No-show
Show     59175    11264
No-show   1126    16692
Random Forest Classifier Classification Report:
              precision    recall  f1-score   support

        Show       0.98      0.84      0.91     70439
     No-show       0.60      0.94      0.73     17818

    accuracy                           0.86     88257
   macro avg       0.79      0.89      0.82     88257
weighted avg       0.90      0.86      0.87     88257

Random Forest Classifier Average ROC AUC Score: 0.9289
```

Hyper parameter Tuning Using GridSearchCV and Derived Features

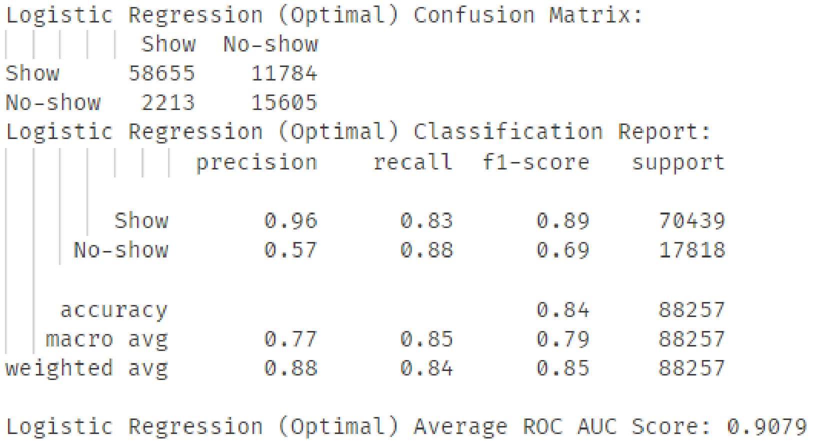
The result with tuning is also greatly improved. With the metric scoring being ROC AUC, the GridSearchCV provided the optimal parameters for Logistic Regression and Random Forest Classifier. However, it is observed the score for Logistic regression was slightly lower with the optimal parameters than the base model. This may be due to the GridSearchCV limitation were implementing random under sampling was not possible. Nevertheless, the ROC AUC score was satisfactory (close to 1). The evaluation of the base and tuned logistic regression models may be done and results compared to see if tuning was impactful.



```
Optimal Parameters and Scores for each model:
lr:
Optimal Parameters: {'C': 1, 'max_iter': 1000, 'penalty': 'l1', 'solver': 'liblinear'}
Score: 0.896095247038778

rf:
Optimal Parameters: {'max_depth': 40, 'min_samples_leaf': 4, 'min_samples_split': 10, 'n_estimators': 150}
Score: 0.9401006893521581
```

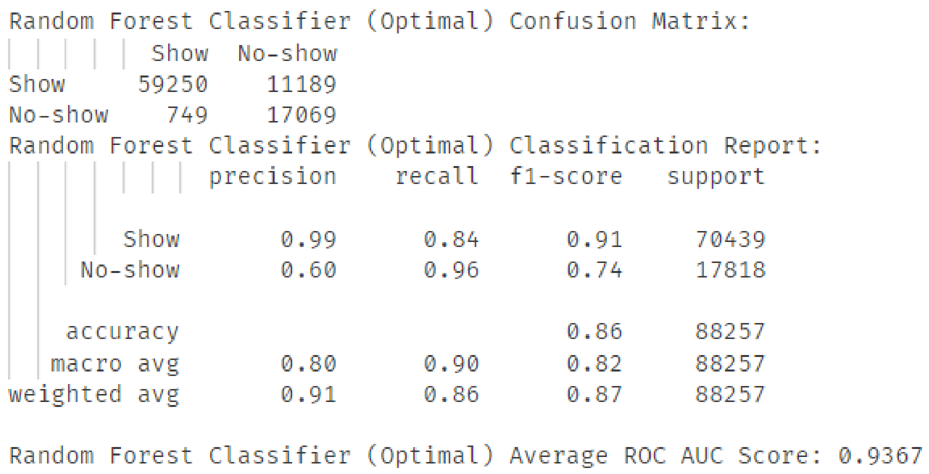
Perform validation using optimal parameters.

To identify the performance difference with and without hyperparameter tuning, validate the models with the parameters set. However, for logistic regression, the difference was negligible (0.9065 - 0.9079) for the ROC AUC score. This means the model did not benefit much from hyperparameter tuning. Instead, feature engineering contributed the most.



```
Logistic Regression (Optimal) Confusion Matrix:
│ │ │ │ │ Show  No-show
Show     58655   11784
No-show   2213   15605
Logistic Regression (Optimal) Classification Report:
│ │ │ │ │ │ │ precision    recall  f1-score   support
│ │ │ │
│ │ │ │   Show       0.96      0.83      0.89     70439
│ │ │ No-show       0.57      0.88      0.69     17818
│ │
│ │ accuracy                           0.84     88257
│ │macro avg       0.77      0.85      0.79     88257
weighted avg       0.88      0.84      0.85     88257

Logistic Regression (Optimal) Average ROC AUC Score: 0.9079
```

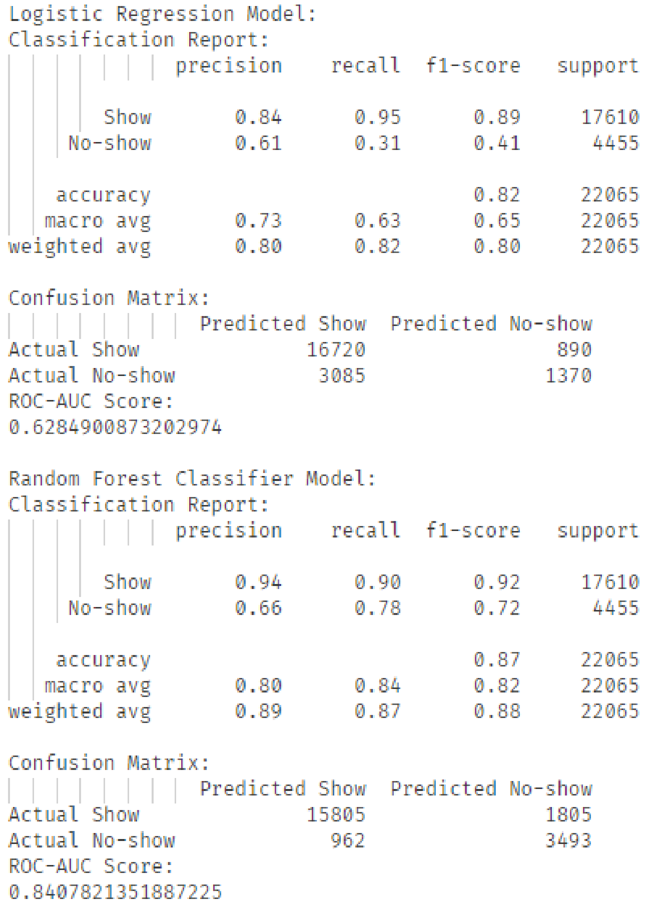
For Random Forest Classifier, the results simply observing the numbers improved with tuning when compared to the base model. It was also higher values than the logistic regression model (base and tuned).



```
Random Forest Classifier (Optimal) Confusion Matrix:
           Show  No-show
Show      59250    11189
No-show     749    17069
Random Forest Classifier (Optimal) Classification Report:
              precision    recall  f1-score   support

        Show       0.99      0.84      0.91     70439
     No-show       0.60      0.96      0.74     17818

    accuracy                           0.86     88257
   macro avg       0.80      0.90      0.82     88257
weighted avg       0.91      0.86      0.87     88257

Random Forest Classifier (Optimal) Average ROC AUC Score: 0.9367
```

Model Testing without Random Under Sampling.

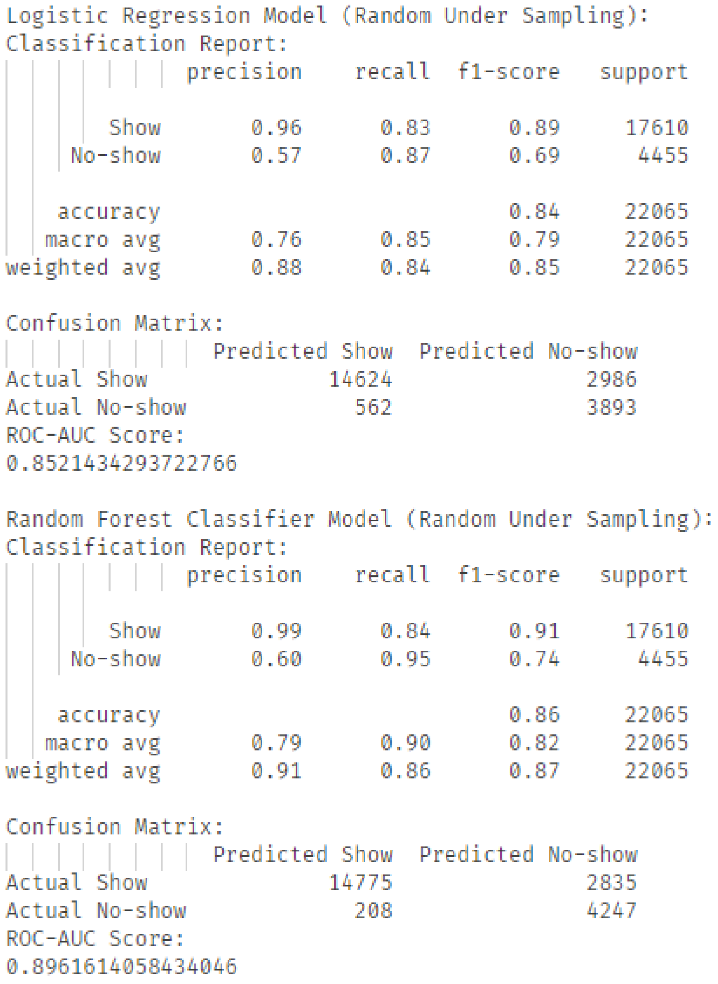
So far, through feature engineering, the model learning process has improved the models' performance significantly as they are able to generalize more on the positive and negative class predictions. Also, hyperparameter tuning the model has been validated and showed a modest but positive impact on the model performance. One can safely proceed to fitting these tuned models on the complete training data and test their performance on the testing set which comprises 20% of the entire dataset. To have exhaustive understanding, this stage performs testing without training data random under sampling and with this resampling. The testing set maintained the class imbalance to mimic the real word as much as possible.



```
Logistic Regression Model:
Classification Report:
              precision    recall  f1-score   support

        Show       0.84      0.95      0.89     17610
     No-show       0.61      0.31      0.41      4455

    accuracy                           0.82     22065
   macro avg       0.73      0.63      0.65     22065
weighted avg       0.80      0.82      0.80     22065

Confusion Matrix:
                Predicted Show  Predicted No-show
Actual Show              16720                890
Actual No-show            3085               1370
ROC-AUC Score:
0.6284900873202974

Random Forest Classifier Model:
Classification Report:
              precision    recall  f1-score   support

        Show       0.94      0.90      0.92     17610
     No-show       0.66      0.78      0.72      4455

    accuracy                           0.87     22065
   macro avg       0.80      0.84      0.82     22065
weighted avg       0.89      0.87      0.88     22065

Confusion Matrix:
                Predicted Show  Predicted No-show
Actual Show              15805               1805
Actual No-show             962               3493
ROC-AUC Score:
0.8407821351887225
```

Model Testing with Training set Random Under Sampling

Randomly under sampling the training set produced high ROC AUC scores for both models. This showed that class imbalance negatively impacts a model's performance. The results are as follows:

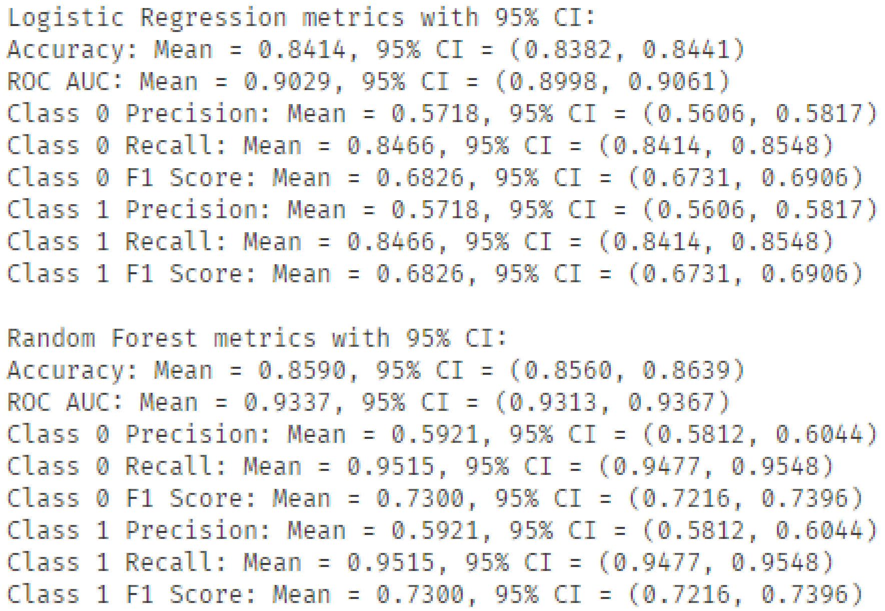


```
Logistic Regression Model (Random Under Sampling):
Classification Report:
              precision    recall  f1-score   support

        Show       0.96      0.83      0.89     17610
     No-show       0.57      0.87      0.69      4455

    accuracy                           0.84     22065
   macro avg       0.76      0.85      0.79     22065
weighted avg       0.88      0.84      0.85     22065

Confusion Matrix:
                Predicted Show  Predicted No-show
Actual Show              14624               2986
Actual No-show             562               3893
ROC-AUC Score:
0.8521434293722766

Random Forest Classifier Model (Random Under Sampling):
Classification Report:
              precision    recall  f1-score   support

        Show       0.99      0.84      0.91     17610
     No-show       0.60      0.95      0.74      4455

    accuracy                           0.86     22065
   macro avg       0.79      0.90      0.82     22065
weighted avg       0.91      0.86      0.87     22065

Confusion Matrix:
                Predicted Show  Predicted No-show
Actual Show              14775               2835
Actual No-show             208               4247
ROC-AUC Score:
0.8961614058434046
```

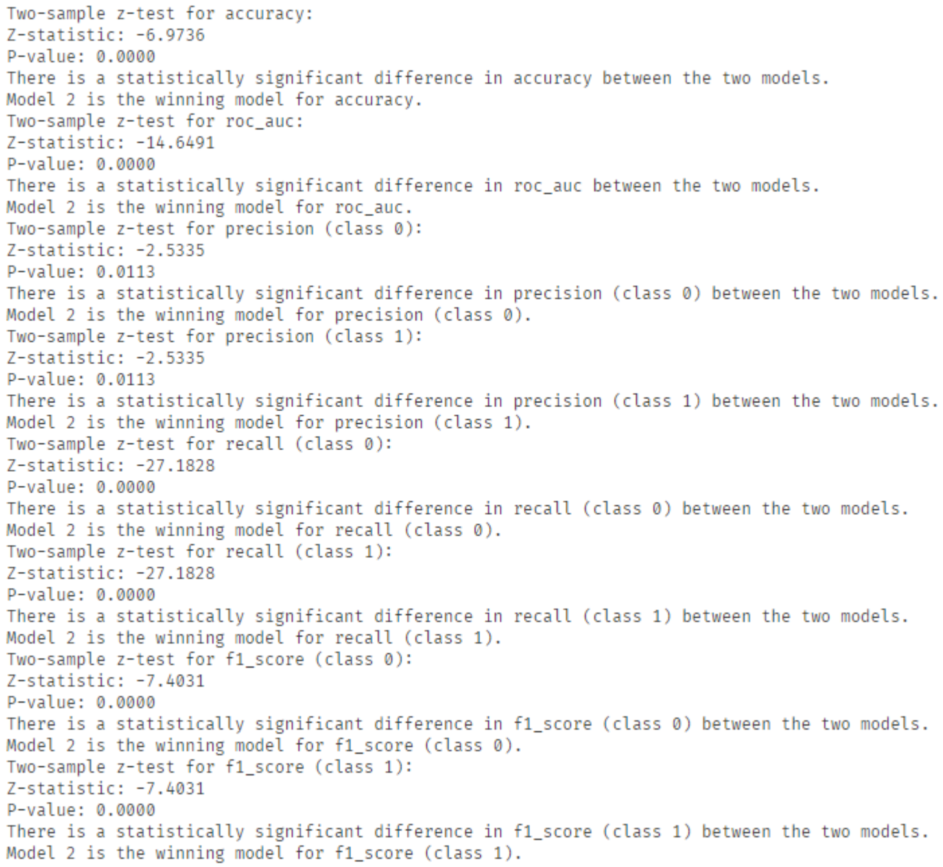
Model Evaluation

In assessing the models' performance on the testing dataset, the confirmation of their ability to predict the No-show class effectively was established. However, moving beyond mere accuracy, which only indicates how many items were classified correctly, a deeper dive into metrics such as ROC AUC score and F1-Score was conducted. These metrics provide a more nuanced understanding of how well the models perform compared to a random guessing classifier. The evaluation revealed that both the Logistic Regression and Random Forest models performed admirably across these metrics. However, to truly discern the significance of the performance differences between these models, statistical tests were conducted at a 95% confidence interval.

For Logistic Regression, a mean accuracy, of 0.8414 (95% CI = 0.8382, 0.8441) and an ROC AUC mean of 0.9029 (95% CI = 0.8998, 0.9061) were found. On the other hand, the Random Forest model had an accuracy mean of 0.8590 (95% CI = 0.8560, 0.8639) and an ROC AUC mean of 0.9337 (95% CI = 0.9313, 0.9367). The statistical analysis using two-sample z-tests unveiled significant differences between the models across all evaluated metrics, including accuracy, ROC AUC, precision, recall, and F1-Score. Interestingly, Model 2 (Random Forest) emerged as the better model, underscoring its effectiveness in predicting appointment no-shows.



```
Logistic Regression metrics with 95% CI:
Accuracy: Mean = 0.8414, 95% CI = (0.8382, 0.8441)
ROC AUC: Mean = 0.9029, 95% CI = (0.8998, 0.9061)
Class 0 Precision: Mean = 0.5718, 95% CI = (0.5606, 0.5817)
Class 0 Recall: Mean = 0.8466, 95% CI = (0.8414, 0.8548)
Class 0 F1 Score: Mean = 0.6826, 95% CI = (0.6731, 0.6906)
Class 1 Precision: Mean = 0.5718, 95% CI = (0.5606, 0.5817)
Class 1 Recall: Mean = 0.8466, 95% CI = (0.8414, 0.8548)
Class 1 F1 Score: Mean = 0.6826, 95% CI = (0.6731, 0.6906)

Random Forest metrics with 95% CI:
Accuracy: Mean = 0.8590, 95% CI = (0.8560, 0.8639)
ROC AUC: Mean = 0.9337, 95% CI = (0.9313, 0.9367)
Class 0 Precision: Mean = 0.5921, 95% CI = (0.5812, 0.6044)
Class 0 Recall: Mean = 0.9515, 95% CI = (0.9477, 0.9548)
Class 0 F1 Score: Mean = 0.7300, 95% CI = (0.7216, 0.7396)
Class 1 Precision: Mean = 0.5921, 95% CI = (0.5812, 0.6044)
Class 1 Recall: Mean = 0.9515, 95% CI = (0.9477, 0.9548)
Class 1 F1 Score: Mean = 0.7300, 95% CI = (0.7216, 0.7396)
```



```
Two-sample z-test for accuracy:
Z-statistic: -6.9736
P-value: 0.0000
There is a statistically significant difference in accuracy between the two models.
Model 2 is the winning model for accuracy.
Two-sample z-test for roc_auc:
Z-statistic: -14.6491
P-value: 0.0000
There is a statistically significant difference in roc_auc between the two models.
Model 2 is the winning model for roc_auc.
Two-sample z-test for precision (class 0):
Z-statistic: -2.5335
P-value: 0.0113
There is a statistically significant difference in precision (class 0) between the two models.
Model 2 is the winning model for precision (class 0).
Two-sample z-test for precision (class 1):
Z-statistic: -2.5335
P-value: 0.0113
There is a statistically significant difference in precision (class 1) between the two models.
Model 2 is the winning model for precision (class 1).
Two-sample z-test for recall (class 0):
Z-statistic: -27.1828
P-value: 0.0000
There is a statistically significant difference in recall (class 0) between the two models.
Model 2 is the winning model for recall (class 0).
Two-sample z-test for recall (class 1):
Z-statistic: -27.1828
P-value: 0.0000
There is a statistically significant difference in recall (class 1) between the two models.
Model 2 is the winning model for recall (class 1).
Two-sample z-test for f1_score (class 0):
Z-statistic: -7.4031
P-value: 0.0000
There is a statistically significant difference in f1_score (class 0) between the two models.
Model 2 is the winning model for f1_score (class 0).
Two-sample z-test for f1_score (class 1):
Z-statistic: -7.4031
P-value: 0.0000
There is a statistically significant difference in f1_score (class 1) between the two models.
Model 2 is the winning model for f1_score (class 1).
```

Conclusion

In conclusion, this project aimed to tackle the issue of missed medical appointments through predictive modeling techniques. By analyzing factors such as SMS reminders, appointment scheduling gaps, and patient demographics, there were uncovered insights into appointment attendance behavior. Logistic regression and random forest classifiers were employed to predict appointment no-shows, with both models demonstrating commendable performance across key metrics such as accuracy, ROC AUC score, precision, recall, and F1-Score. However, feature engineering was key to achieving these performance metrics. Class imbalance was demonstrated to impede the model's performance, with the model learning more on predicting the majority which could be useful in some cases but generally not desired. Statistical tests conducted at a 95% confidence interval also showed some notable differences between the models, with the random forest model inferred as the better model for predicting no shows.

The findings highlighted the potential of data mining and machine learning in healthcare settings, offering opportunities to optimize allocating resources for improving patient's attendance. By leveraging predictive models informed by patient data, healthcare institutions can implement targeted interventions to improve appointment attendance rates and mitigate the challenges posed by missed appointments. The project's objective was completed to an acceptable degree of model performance, quality, and data understanding. Moving forward, further research and implementation of predictive modeling techniques hold promise for advancing appointment management strategies and ultimately improving patient outcomes.