36103 Statistical Thinking for Data Science

Assessment 3 | Modelling & Deriving Insights

Nathan Collins

12062131

Assessment 3:	**Modelling and Deriving Insights**
Type:	Individual Assessment
Deliverables:	Jupyter Notebook (x1)
	Final Report **1098 words**
Weight:	50%
Due:	Sunday, 5th November, 23:59

Assessment Criteria:

- Clarity in articulating the questions along with a well-defined proposal for making the invisible visible for a specified set of stakeholders.
- The soundness of the statistical methodology also shows evidence of having applied relevant analytical methods to address the business questions.
- Appropriateness of the interpretation applied to the results, where the resulting conclusions are well justified and answer the business questions.
- Clarity and fluency in communicating your findings to a technical target audience. The soundness of the model interpretation and implications, and professionalism of the executive summary for decision makers.

Section 1 Business Understanding

[1.1] Business Objective & Data Mining Goals

A telecommunications company has launched a new marketing campaign, promoting a subscription plan to their customers. The company seeks to identify customer segments that indicate high responsiveness to their **recent** campaign.

To achieve this, the construction of statistical learning models to gauge its success will be carried out, followed by an evaluation of the performance of each model. Once conducted, insights will be extrapolated to assist strategic decisions for future marketing campaigns.

[1.2] Hypothesis

Null | *The **new** marketing campaign has **no significant** influence on the subscription uptake among customers and there are no customer segments that display high sensitivity to the marketing campaigns.*

Alternate | *The **new** marketing campaign has **a significant** influence on the uptake of the subscription plan among customers and there are customer segments that display high sensitivity to the marketing campaigns.*

Ethical Considerations

*The misuse of data can lead to ethical consequences, even when limited personal information is involved; because of this, the potential outcomes fashioned by third parties must always be considered. These consequences could manifest as violations of individuals' **privacy** within the dataset or the emergence of **biases** in decision-making resulting from its use. Ensuring **unauthorised access** doesn't take place following deployment must also be considered by implementing robust **security** protocols to mitigate data breaches.*

Section 2 Data Understanding and Preparation

[2.1] Understanding the Data

The dataset comprises **41180** observations across **21** variables.

Observations entail separate customers and features associated with that customer.

These include the **target** outcome "**y**" (where **yes** indicated **subscribed**), alongside other noteworthy traits, such as

- **age**,
- **education** background,
- contacts performed during a **campaign**,
- and **employment variation rate**.

< See the appendix for a complete list of the features. >

By consulting the data dictionary, the dataset illustrates features concerning the prior campaign too – where the feature "**previous**" refers to the number of contacts made before the campaign of interest, and "**poutcome**", is the outcome of the previous campaign. This indicates that there is a cohort within the dataset who were not exposed to the previous campaign.

This will need to be considered prior to modelling.

[2.2] Data Preparation

Prior to analysis, data was organised by applying conventional cleaning and manipulation techniques; first by transformation into a pandas data frame, followed by specific feature conversion into integer values.

< For a full recount of descriptions pertaining to the pre-processing steps, see the Exploratory Analysis Report. >

Duplicate Rows

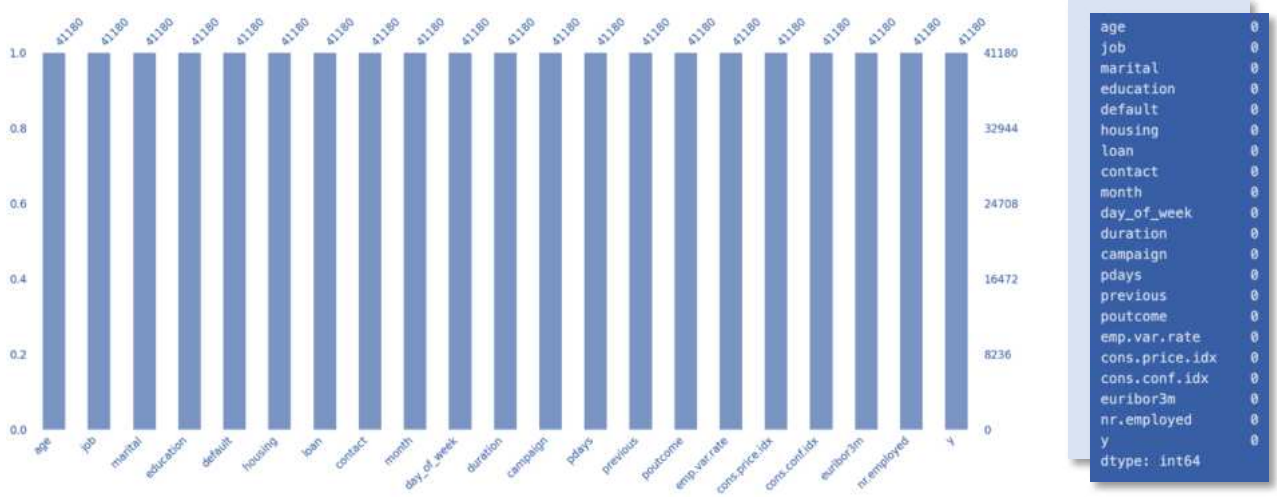
12 duplicate rows were identified and omitted as these outcomes may influence statistical modelling outcomes (Figure 1).

```
Number of duplicated entries: 12
Duplicate entries:
       age          job   marital           education  default housing loan  \
1262    39  blue-collar   married            basic.6y       no      no   no
12257   36      retired   married             unknown       no      no   no
14230   27   technician    single  professional.course       no      no   no
16952   47   technician  divorced         high.school       no     yes   no
18461   32   technician    single  professional.course       no     yes   no
20212   55     services   married         high.school  unknown      no   no
20530   41   technician   married  professional.course       no     yes   no
25213   39      admin.   married    university.degree       no      no   no
28473   24     services    single         high.school       no     yes   no
32512   35      admin.   married    university.degree       no     yes   no
36947   45      admin.   married    university.degree       no      no   no
38277   71      retired    single    university.degree       no      no   no
```

Figure 1 *Duplicate rows of the dataset that were identified and omitted.*

Missing or Vague Variables

No “NaN” or missing values were identified (Figure 2), nor were any outliers excluded, as these represented real scenarios. All implausible “**999**” values in **pdays** remained untouched, though future modelling should consider replacing each with 0.




```
age               0
job               0
marital           0
education         0
default           0
housing           0
loan              0
contact           0
month             0
day_of_week       0
duration          0
campaign          0
pdays             0
previous          0
poutcome          0
emp.var.rate      0
cons.price.idx    0
cons.conf.idx     0
euribor3m         0
nr.employed       0
y                 0
dtype: int64
```

Figure 2 *NaN values visualised with Missingno.*

Omission of “unknown” variables

Over half of the rows in the dataset contained at least one entry with an unknown status. As these variables can impede modelling interpretation, all were omitted in a supplementary data frame.

```
Column Name: job, Unknown Count: 330
Column Name: marital, Unknown Count: 80
Column Name: education, Unknown Count: 1730
Column Name: default, Unknown Count: 8595
Column Name: housing, Unknown Count: 990
Column Name: loan, Unknown Count: 990
```

```
Total rows in the dataset with unknown variables:  41168
Total rows in the dataset after dropping unknown variables:  30471
```

Figure 3 *Counting unknown variables in the given dataset (top), total rows where an unknown variable exists and the total rows following removal of all unknown variables (bottom).*

Encoding & Standardising

Categorical features were successively encoded to embody binary outcomes (Figure 4), followed by scaling to avoid inflated importance throughout the modelling phase, (e.g. **pdays**, which bears mostly 999 values).

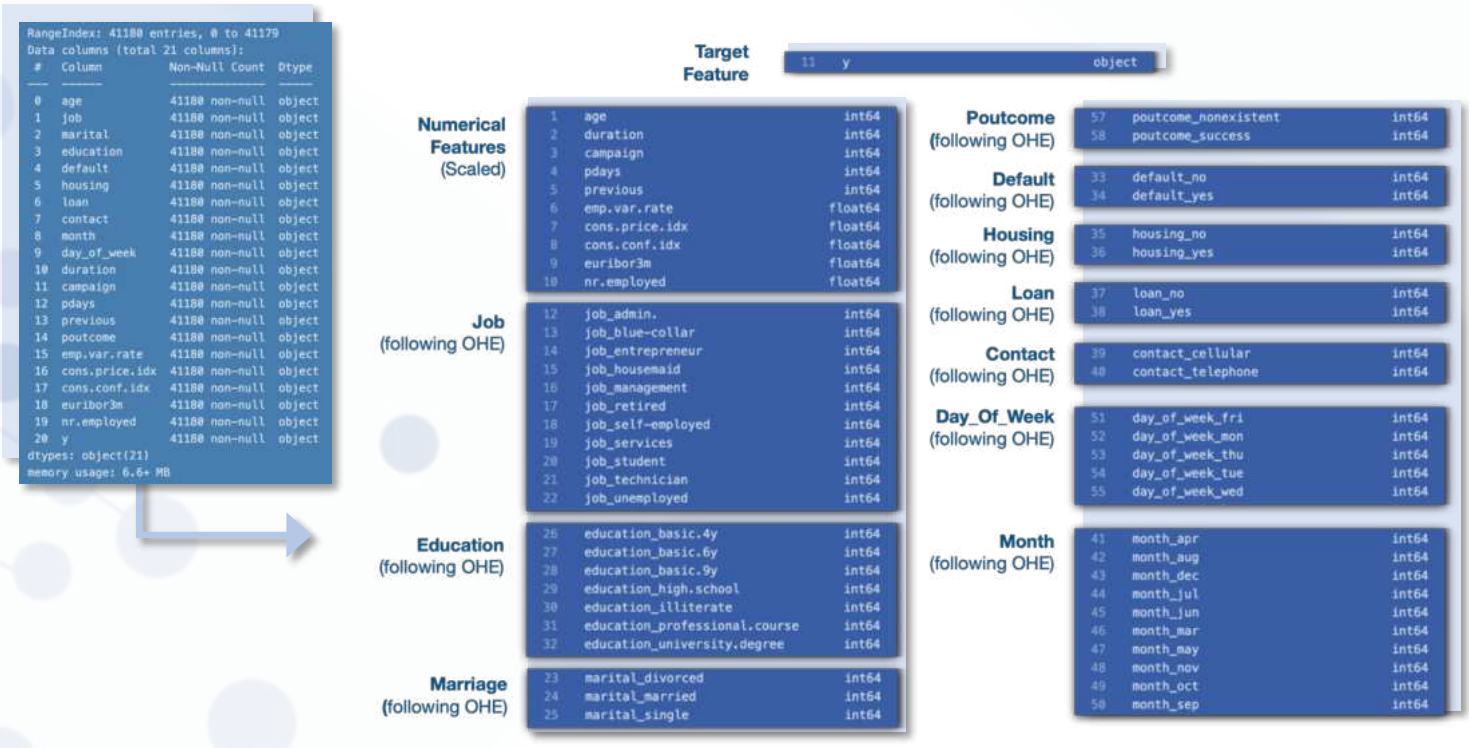


Figure 4 The transformation performed on the provided data frame (left), and the finalised data frame used for modelling (right).

Section 3 Exploratory Analysis Insights

Key Insights:

The dataset is considered **unbalanced**, as the target cohort represents a small minority of both the present and previous campaigns. This may influence the modelling phase (see Figure 5).

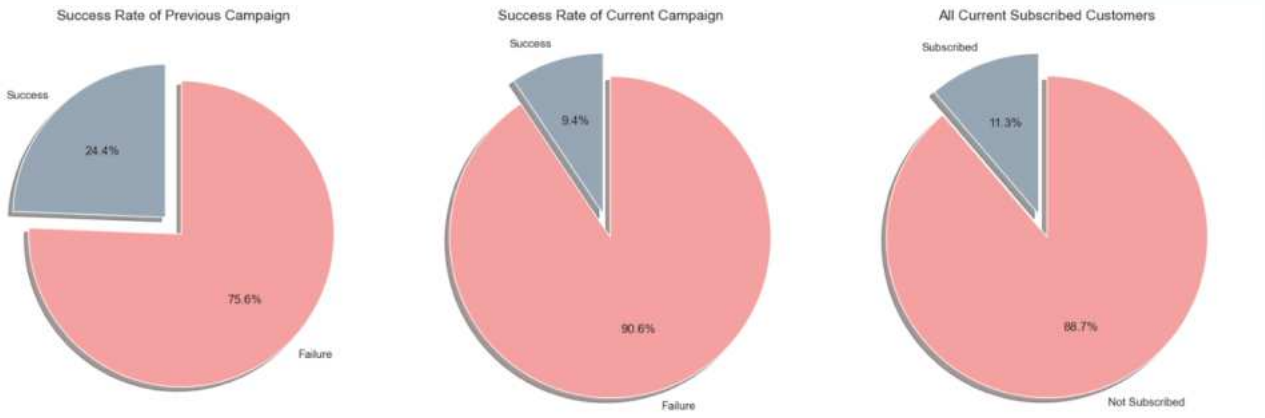


Figure 5 Three pie charts illustrating the success rates of both campaigns, followed by all currently subscribed customers.

Key Insights:

The most sensitive to the new campaign are **married, early 30s and university educated individuals, who have access to housing*, primarily uses cellular and has had at least one contact during the campaign**. This individual is most receptive on a Thursday, at the conclusion of Autumn, and was not exposed to the previous campaign (see Figure 6).

	Previous Campaign	Current Campaign	All Current Subscribers
age	29	33	31
job	admin.	admin.	admin.
marital	married	married	married
education	university.degree	university.degree	university.degree
default	no	no	no
housing	yes	yes	yes
loan	no	no	no
contact	cellular	cellular	cellular
month	may	may	may
day_of_week	thu	thu	thu
duration	192	209	301
campaign	1	1	1
pdays	3	999	999
previous	1	0	0
poutcome	success	nonexistent	nonexistent
emp.var.rate	-1.8	-1.8	-1.8
cons.price.idx	92.893	92.893	92.893
cons.conf.idx	-46.2	-46.2	-46.2
euribor3m	0.879	4.962	4.962
nr.employed	4991.6	5099.1	5099.1
y	yes	yes	yes

Figure 6 A series of data frames comparing the traits of subscribers from the previous campaign, subscribers from the current campaign, and all present subscribers.

Section 4 Modelling

a) The datasets

Experimentation employed an iterative methodology, by trialling & and assessing **16** constructed **variations** of the original dataset.

Key distinctions between each dataset entailed**:**

The original dataset, cleaned.
The dataset, cleaned, with **resampling.** (integrated to mediate the imbalance seen in the target cohort.)
The dataset, cleaned with **unknown outcomes omitted**.
The dataset, cleaned, with integrated **feature engineering**.
The dataset, cleaned, with **duration** omitted. (as this feature isn't available until after the campaign.)

The remaining 11, were hybrids of each of these alterations.

A complete table of each dataset is listed below.

< A list of all engineered features is available in the appendix. >

DataFrame	Description
df	The original, cleaned dataframe.
df_no_duration	The original, cleaned dataframe without the "duration" feature.
df_no_unknowns	The original, cleaned dataframe (with unknown values omitted).
df_no_unknowns_no_duration	The original, cleaned dataframe with unknown values omitted and without the "duration" feature.
df_resampled	The original, cleaned dataframe (with resampling to account for the unbalanced dataset).
df_resampled_no_duration	The original, cleaned dataframe with resampling to account for the unbalanced dataset and without the "duration" feature.
df_no_unknowns_resampled	The original, cleaned dataframe (with unknown values omitted and resampling to account for the unbalanced dataset).
df_no_unknowns_resampled_no_duration	The original, cleaned dataframe with unknown values omitted, resampling for the unbalanced dataset, and without the "duration" feature.
df_feature_engineering	The original, cleaned dataframe (with feature engineering applied).
df_feature_engineering_no_duration	The original, cleaned dataframe with feature engineering applied and without the "duration" feature.
df_no_unknowns_feature_engineering	The original, cleaned dataframe (with unknown values omitted and feature engineering applied).
df_no_unknowns_feature_engineering_no_duration	The original, cleaned dataframe with unknown values omitted, feature engineering applied, and without the "duration" feature.
df_resampled_feature_engineering	The original, cleaned dataframe (with resampling to account for the unbalanced dataset and feature engineering applied).
df_resampled_feature_engineering_no_duration	The original, cleaned dataframe with resampling for the unbalanced dataset, feature engineering applied, and without the "duration" feature.
df_no_unknowns_resampled_feature_engineering	The original, cleaned dataframe (with unknown values omitted, resampling to account for the unbalanced dataset and feature engineering applied).
df_no_unknowns_resampled_feature_engineering_no_duration	The original, cleaned dataframe with unknown values omitted, resampling for the unbalanced dataset, feature engineering applied, and without the "duration" feature.

Figure 7 A complete list of the dataframes modelled.

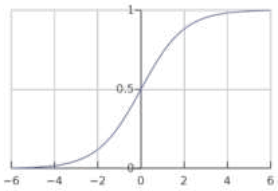
b) Modelling

As a statistical model can be leveraged for predictive insights that may augment stakeholder decisions and strategy, the project's modelling phase considered parametric and non-parametric approaches.

Models Applied

Parametric Modelling

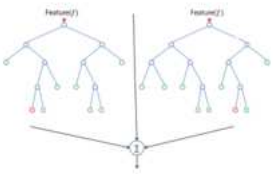
Logistic Regression | **LR**



LR was the **parametric** model choice. While considered straightforward, it serves as a reliable comparative staple, offering clarity for interpretability. It's particularly effective when relationships between features and the response variable resemble a linear pattern, as this may provide our stakeholders with a direct understanding of specific feature influences.

Non-Parametric Modelling

Random Forest | **RFC**



For the **non-parametric** approach, an **RFC** was selected. This model aggregates results from multiple decision trees to generate predictions, thereby ensuring a more robust and possibly non-linear understanding of the dataset. Notably, it also offers an assessment of feature importance, which can be invaluable for strategic insights.

Supplementary Non-Parametric Modelling

Extreme Gradient Boost | **XGB**

Recognized for its speed, performance, and in-built regularisation techniques, **XGB** was chosen to potentially enhance model accuracy and mitigate overfitting through its wide variety of hyperparameters.

K-Nearest Neighbour | **KNN**

KNN was employed, as it's direct and provides no underlying assumptions about data distributions. KNN models offer versatility and are applicable in broad contexts. (Outcomes available in the appendix).

- Datasets were each modelled in **sequence** and were **aggregated** for interpretation.
- Each iteration explored **multiple variations** of the dataset representing the same cohort.
- Non-parametric models were each subjected to **hyperparameter** tuning and **regularisation**.

Maximum Likelihood Estimation (MLE)

As prior knowledge about the nature of collection was limited, an MLE approach was advised over Bayesian methods. MLE's strength is in determining model parameters without relying on prior distributions, and given its adaptability and efficiency, it served as an appropriate choice in this context, ensuring our analysis remained unbiased and grounded in the observed data.

As we recognise that terminology may be unfamiliar to our stakeholders, our team remains available for clarification or explanation about specific models and insights.

Section 5 Results

A Guide to Interpreting the Performance Metrics

Accuracy | measures the proportion of **true results** (both true positives and true negatives) among the total number of cases examined.

AUC (Area Under the Curve) | represents the degree to which a model is capable of **distinguishing between classes;** higher AUCs indicate a better-performing model.

Error Rate | the proportion of all **incorrect predictions** out of the total predictions made, (one minus the accuracy).

Sensitivity (True Positive Rate) | measures the proportion of **actual positives** correctly identified, also known as "recall" in some contexts.

Specificity (True Negative Rate) | measures the proportion of **actual negatives** that are correctly identified.

Precision | quantifies the number of **true positives** out of all the positive results predicted by the model.

Recall (Sensitivity) | reflects the **fraction of positives** that were correctly identified by the model.

Log Loss | an indication of the **accuracy** of a classifier by penalizing false classifications; lower log loss indicates a more accurate model.

The Baseline

To gauge the influence of each model, a simple baseline was established prior to experimentation. The following is set as a point of reference for each experiment and iteration. **Without resampling**, baseline accuracy resided at 89%.

DataFrame	Baseline Accuracy
df	88.74%
df_no_duration	88.74%
df_no_unknowns	87.35%
df_no_unknowns_no_duration	87.35%
df_resampled	50.00%
df_resampled_no_duration	50.00%
df_no_unknowns_resampled	50.00%
df_no_unknowns_resampled_no_duration	50.00%
df_feature_engineering	88.74%
df_feature_engineering_no_duration	88.74%
df_no_unknowns_feature_engineering	87.35%
df_no_unknowns_feature_engineering_no_duration	87.35%
df_resampled_feature_engineering	50.00%
df_resampled_feature_engineering_no_duration	50.00%
df_no_unknowns_resampled_feature_engineering	50.00%
df_no_unknowns_resampled_feature_engineering_no_duration	50.00%

Figure 8 A complete list of each dataframes baseline level of accuracy.

While concise details are reported, an extensive array of visuals and tables covering all 16 datasets are provided in the report appendix.

Logistic Regression | **LR**

While **duration** appeared as one of the more influential features, it was not known until after the campaign. Following omission, model accuracies decline considerably.

The **resampled** models offer a high **AUC**, though can be improved (see Section 6).

DataFrame	Accuracy	AUC	Error Rate	Sensitivity	Specificity	Precision	Recall	Log Loss
df_resampled	0.949702	0.991663	0.050298	0.926098	0.973310	0.971991	0.926098	0.120398
df_feature_engineering	0.914744	0.939931	0.085256	0.443231	0.973763	0.678930	0.443231	0.202720
df	0.915837	0.939617	0.084163	0.443231	0.974993	0.689304	0.443231	0.202501
df_no_unknowns	0.908942	0.935694	0.091058	0.443828	0.971344	0.675105	0.443828	0.213096
df_no_duration	0.903206	0.786583	0.096794	0.241266	0.986062	0.684211	0.241266	0.275317

Figure 10 Testing set: Five of the most distinct data frame variations, and their corresponding performance metric, following a hpyerparamter-tuned Logistic Regression (see appendix for ranges).

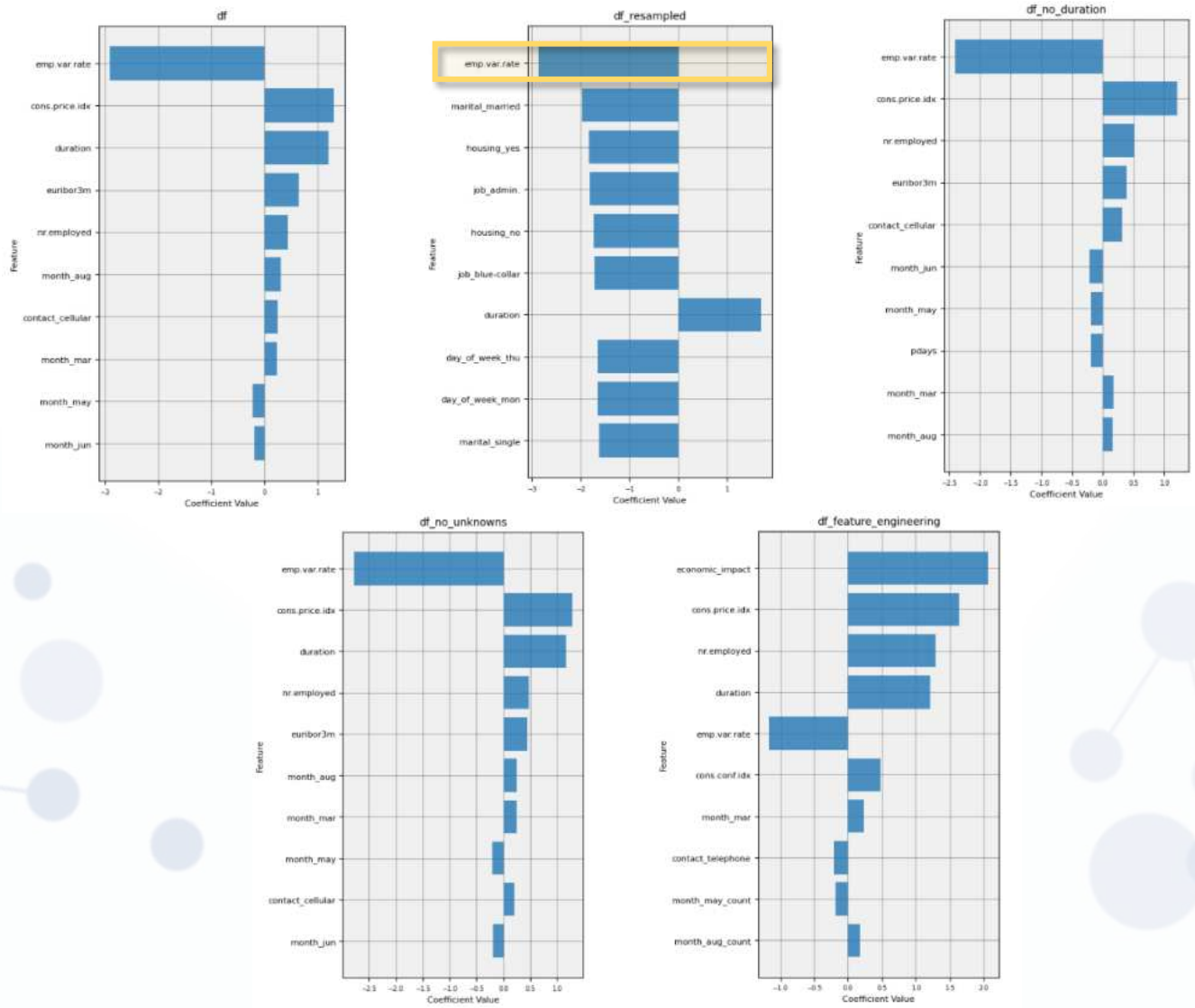


Figure 11 The Coefficient values denoting the features of strongest influence over each data frame.

Logistic Regression | **Receiver Operating Characteristics (ROC) Curves**

*An **ROC curve** is a chart that denotes how well a test can separate two things, this can be thought of as comparing "sick people from healthy people", by measuring the balance between catching true cases and avoiding false alarms. The better the test, the closer the curve will be to the top-left corner of the chart. It is worth recognising the balanced or "resampled" dataframes provide more improved ROC curves, than other data frames, meaning a balanced dataset increases its accuracy.*

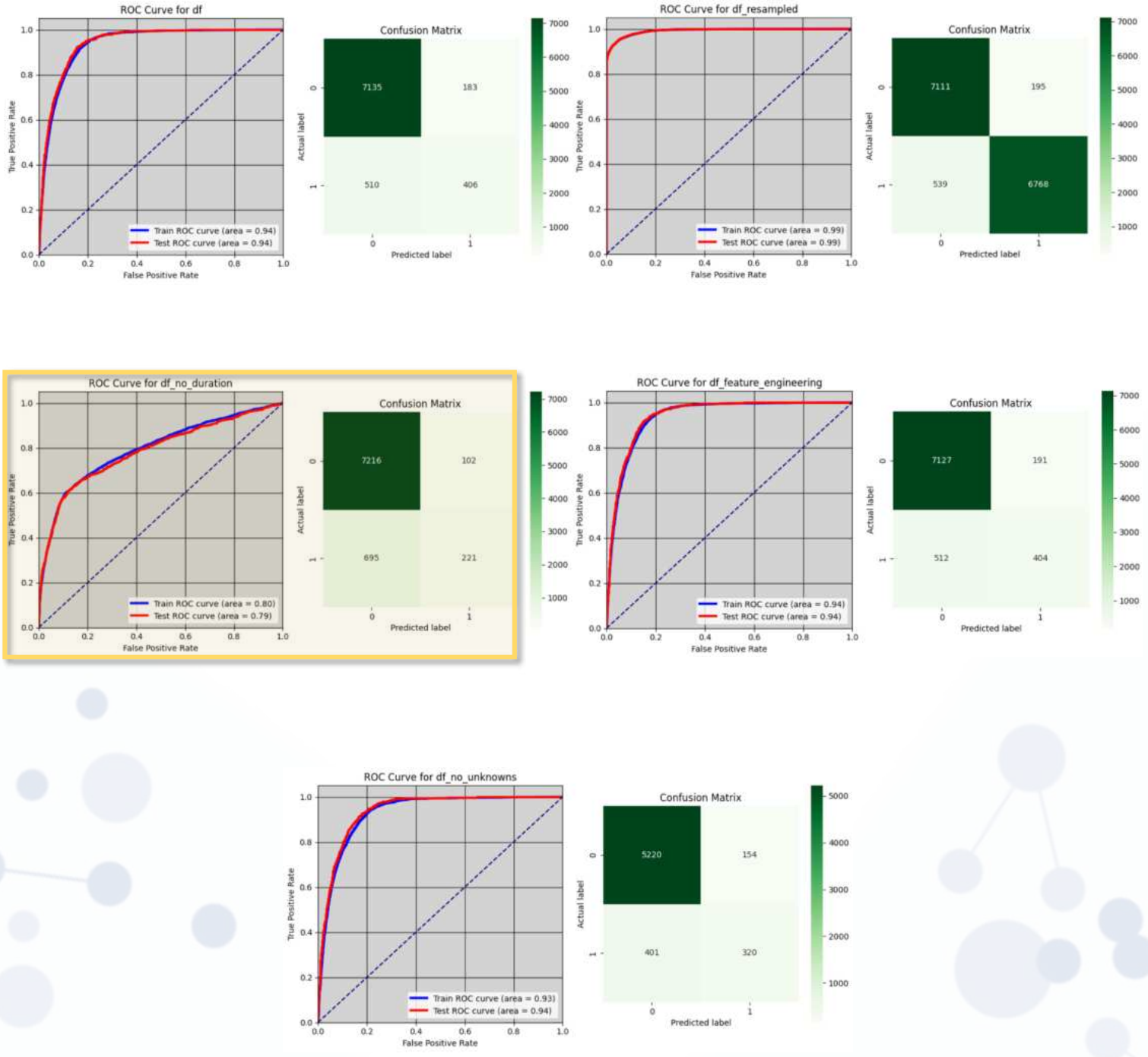


Figure 12 Five ROC curves and corresponding confusion matrixes, denoting the true positives and true negative success rates, for the Logistic Regression modelling phase.

Random Forest Classifier | **RFC**

DataFrame	Accuracy	AUC	Error Rate	Sensitivity	Specificity	Precision	Recall	Log Loss
df_resampled	0.953808	0.993043	0.046192	0.945395	0.962223	0.961581	0.945395	0.115834
df_no_unknowns	0.909598	0.943183	0.090402	0.475728	0.967808	0.664729	0.475728	0.190572
df	0.912558	0.938911	0.087442	0.465066	0.968571	0.649390	0.465066	0.208332
df_feature_engineering	0.911708	0.938435	0.088292	0.453057	0.969117	0.647426	0.453057	0.202675
df_no_duration	0.891912	0.765718	0.108088	0.280568	0.968434	0.526639	0.280568	0.496763

Figure 13 Testing dataset: Five of the most distinct data frame variations, and their corresponding performance metrics for a Random Forest Classifier, following hyperparameter tuning (see appendix for ranges).

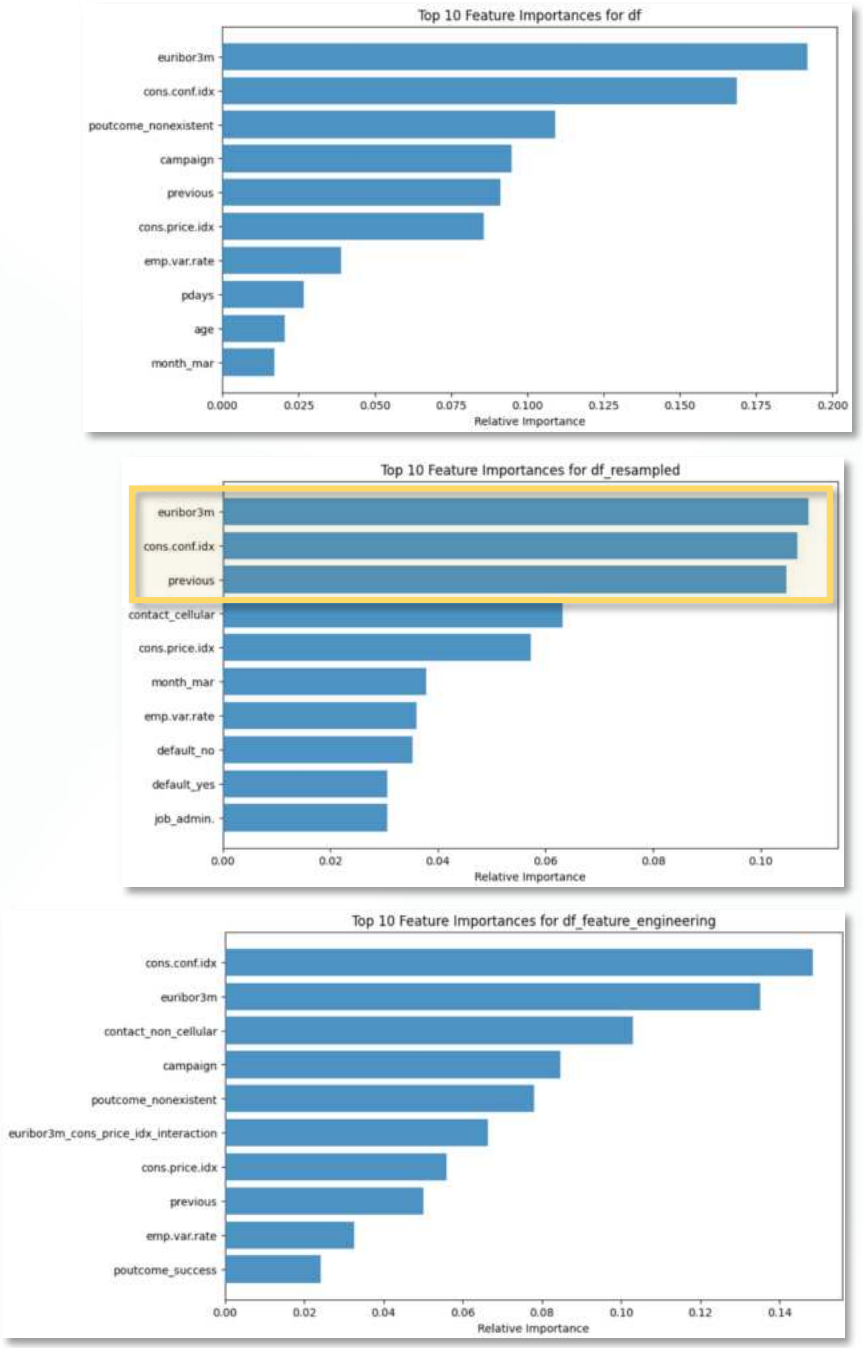


Figure 14 The Feature Importances denoting the columns of strongest influence within each data frame. The three with the most variation in data frame structure were selected.

Random Forest Classifier | **ROC Curves**

*Despite tuning models with the best-performing hyperparameters, there remained a considerable **gap** in the **seen** vs **unseen data**. Although the random forest has a greater predictive capacity for known data, it is considered to perform worse than the logistic regression. It is worth acknowledging the considerable gap in seen vs unseen data, following the omission of the **duration** feature (see Figure 3 below).*

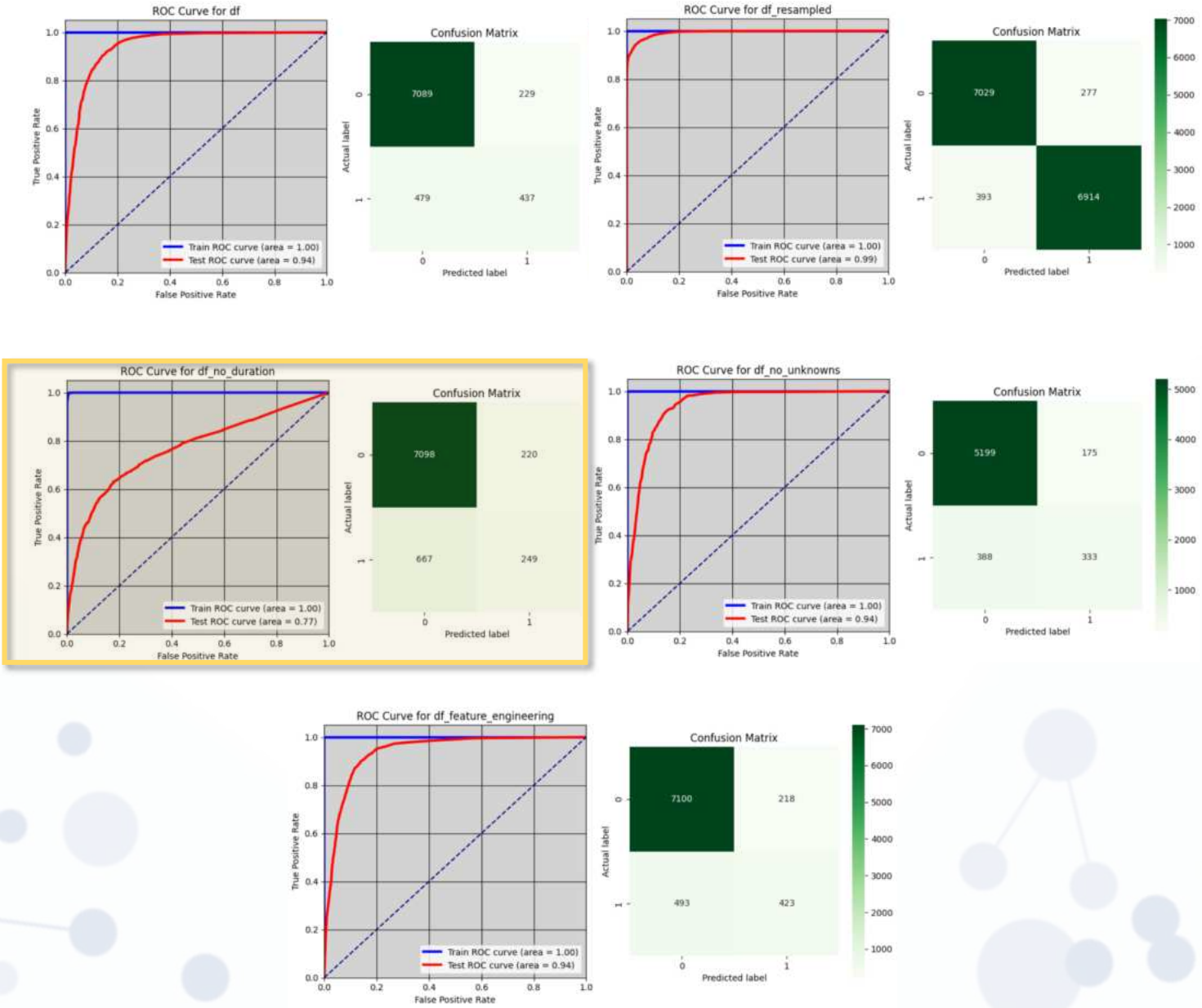
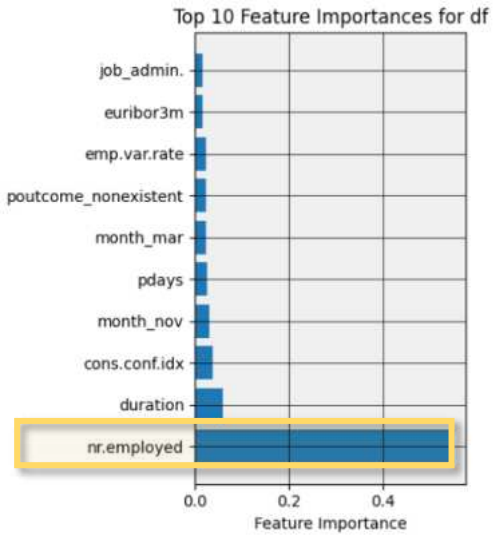


Figure 15 Five ROC curves and corresponding confusion matrixes, denoting the true positives and true negative success rates, for the Random Forest Classification modelling phase.

Extreme Gradient Boost | **XGB**

DataFrame	Accuracy	AUC	Error Rate	Sensitivity	Specificity	Precision	Recall	Log Loss
df_resampled	0.949429	0.992798	0.050571	0.944437	0.954421	0.953967	0.944437	0.108997
df_feature_engineering	0.917537	0.949971	0.082463	0.550218	0.963515	0.653696	0.550218	0.170488
df_no_unknowns	0.917637	0.951113	0.082363	0.571429	0.964086	0.680992	0.571429	0.174702
df	0.919480	0.950672	0.080520	0.555677	0.965018	0.665359	0.555677	0.169603
df_no_duration	0.903328	0.798267	0.096672	0.255459	0.984422	0.672414	0.255459	0.270559

Figure 16 Testing dataset: Five of the most distinct data frame variations, and their corresponding performance metrics for a XGBoost Classifier, following hyperparameter tuning (see appendix for ranges).



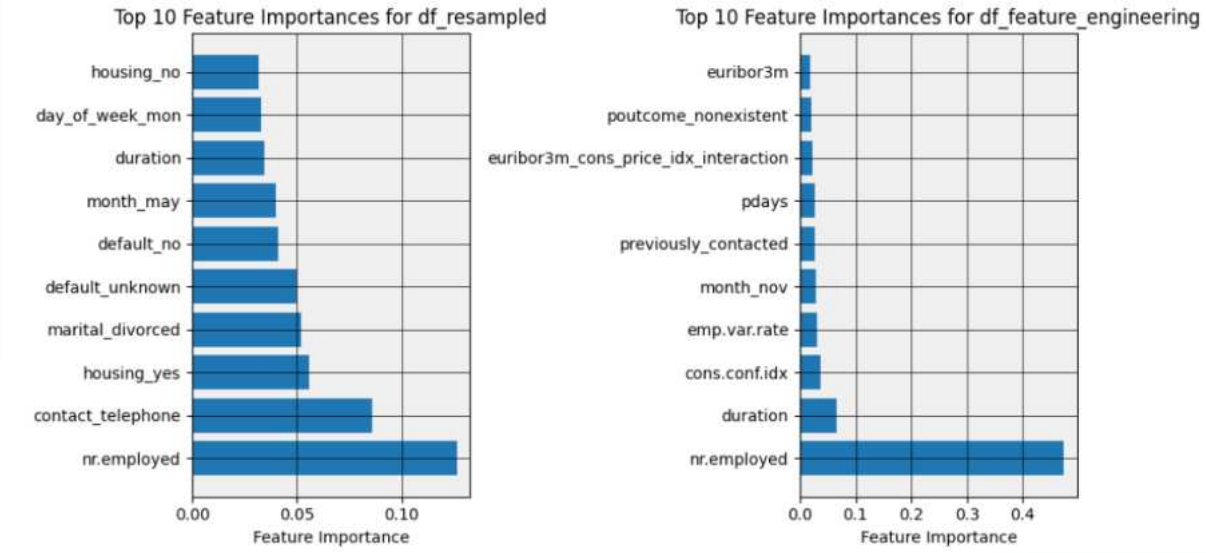


Figure 17 The Feature Importances denoting the columns of strongest influence within each data frame. The three with the most variation in data frame structure were selected.

XGBoost Classifier | **ROC Curves**

*The XGBoost classifier shares a similar testing df_no_duration as the Logistic regression model. This model, however, is considered **less accurate** than the logistic regression, as the training and testing ROC curves denote a wider variance (0.12, vs 0.01).*

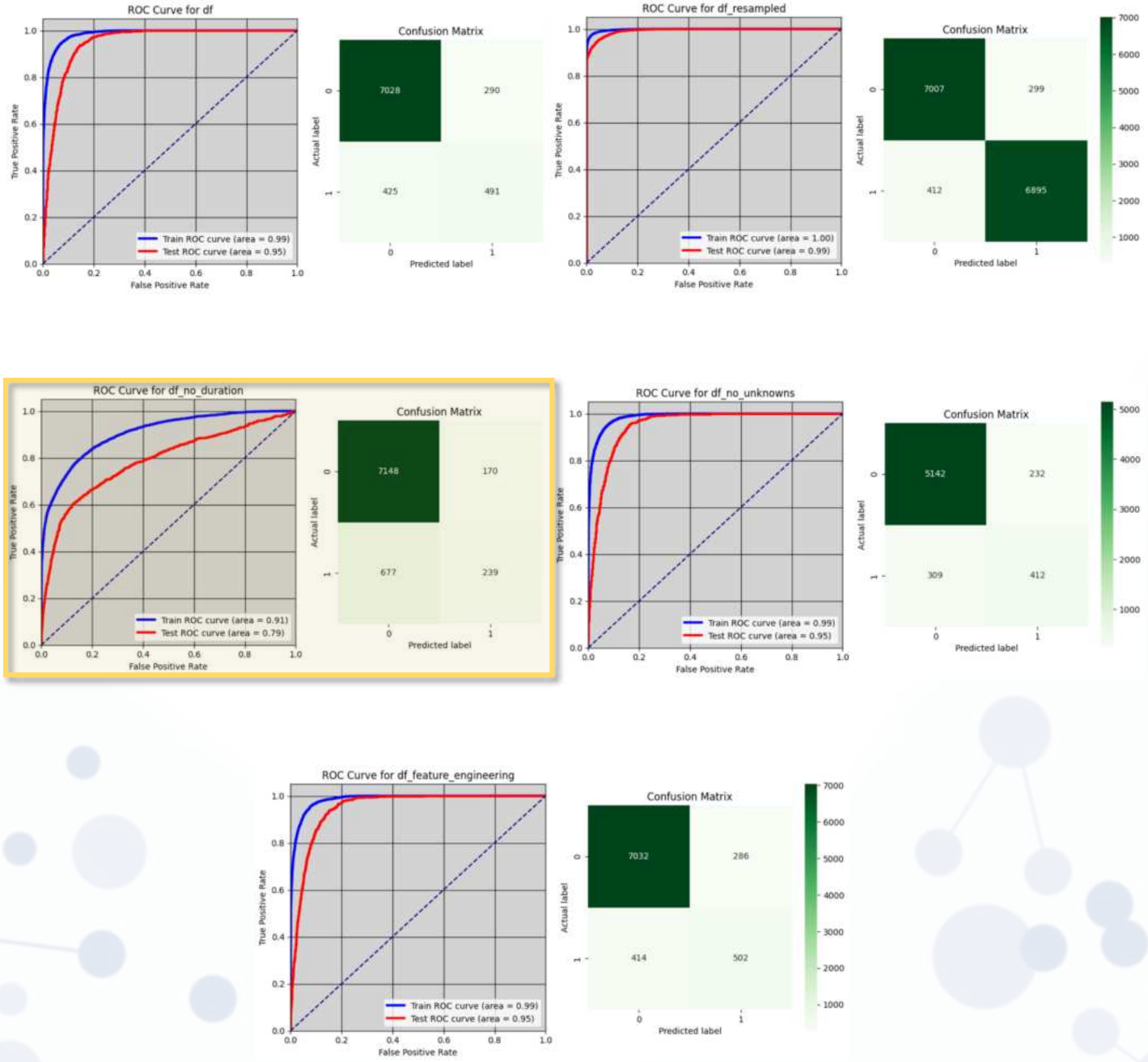


Figure 18 Five ROC curves and corresponding confusion matrixes, denoting the true positives and true negative success rates, for the XGBoost Classification modelling phase.

Section 6 Stakeholder Insights

Logistic Regression

Financial indicators of existing customers are significant markers of positive reception to the latest campaign, most notably:

emp.var.rate and **cons.price.indx**.
(-1.8) (92.9)

Following resampling,

personal features, such as **marital status** and **housing**, rise in influence (Figure 19, 97% ROC).

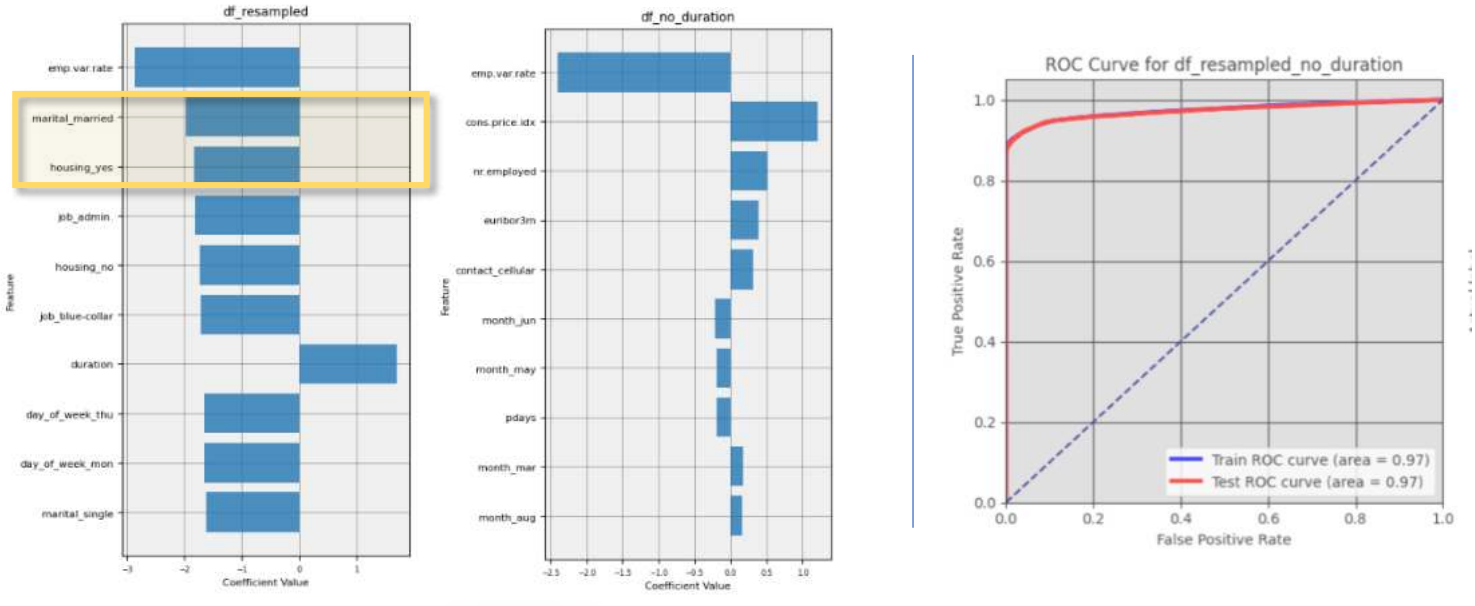


Figure 19 The most influential features following Logistic Regression, comparing the data frame without "duration" and "resampling".

Stakeholders are advised to prioritise economic and employment factors surrounding a customer's purchasing habits, with a specific emphasis on the quarter employment variation rate. It is also advised to accumulate features associated with these factors, such as their "**duration within a particular occupation**", their respective "**activity**" in the occupation, their "**salary**", or their "**residential suburb**" prior to future campaigns.

Random Forest Classifier

While the strongest performing model is in the seen (training) data, RFC is weaker than the previous models when applied to unseen (testing) data. This is apparent, especially when omitting rows with **unknown** values and the **duration** feature (see Figure 21).

While the RFC is considered not as refined as the logistic regression, it reinforces the value of occupational and finance factors that influence customer reception of the latest campaign. In this circumstance, it relates to **euribor3m** (Euribor 3-month rate, 4.96) and once more, **cons.price.indx.**

A customer's response to the **previous** campaign is also likely to see a positive response to the succeeding campaign.

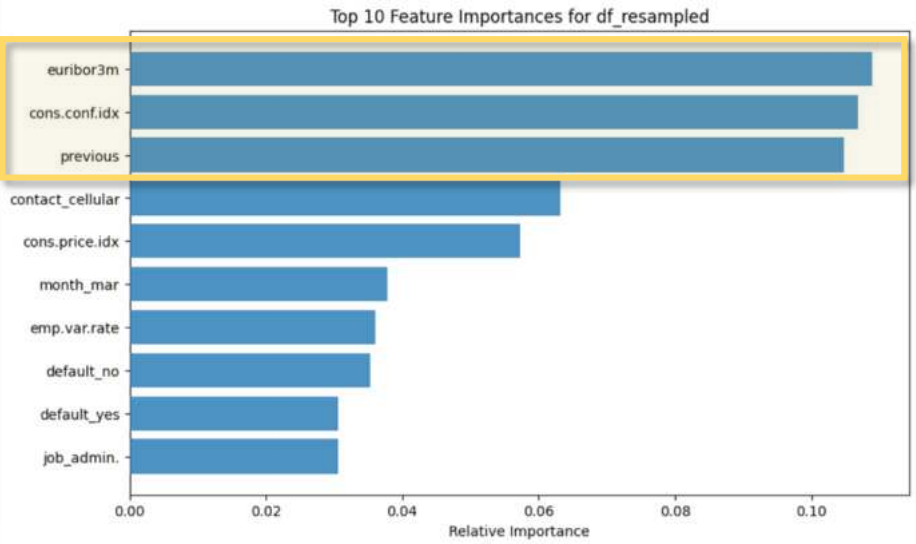


Figure 20 The most influential features following resampling prior to a Random Forest Classification.

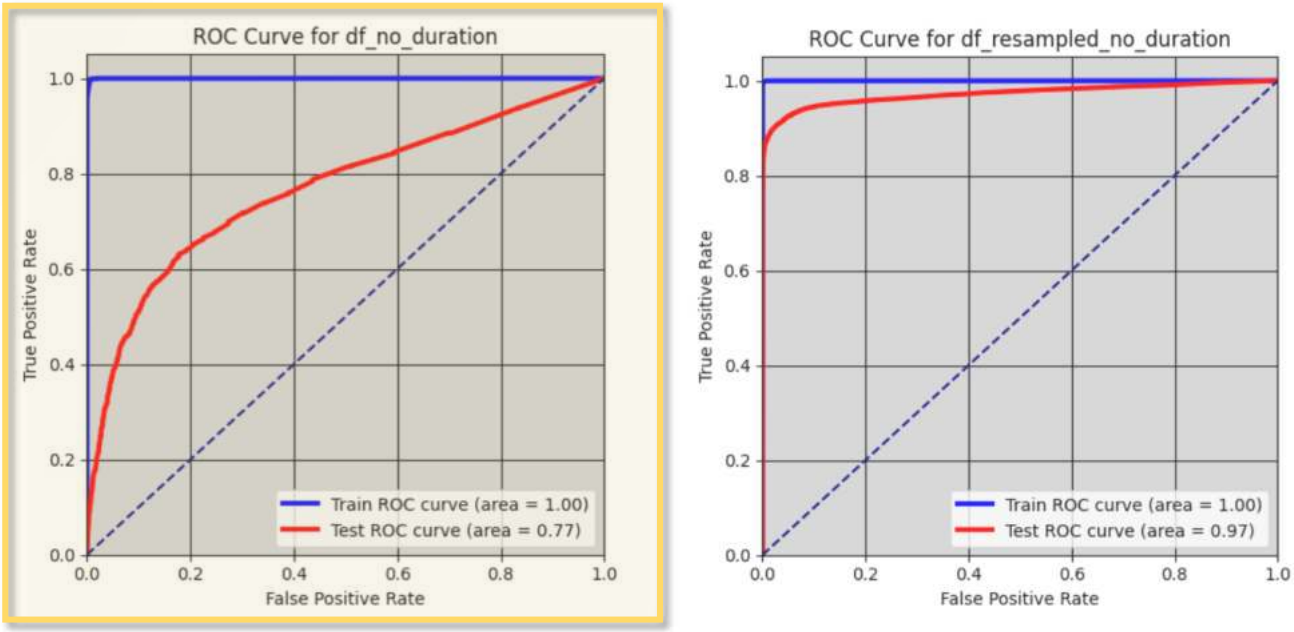


Figure 21 Comparing the model's ROC against seen (training) data, vs unseen (testing) data.

XGBoost Classifier

While also sharing a similar ROC gap between the training and testing datasets, the XGBoost classifier once more reinforces occupational-derived features, emphasising **nr.employed** (5099) as its most important feature of both the **cleaned** and **resampled** data frames.

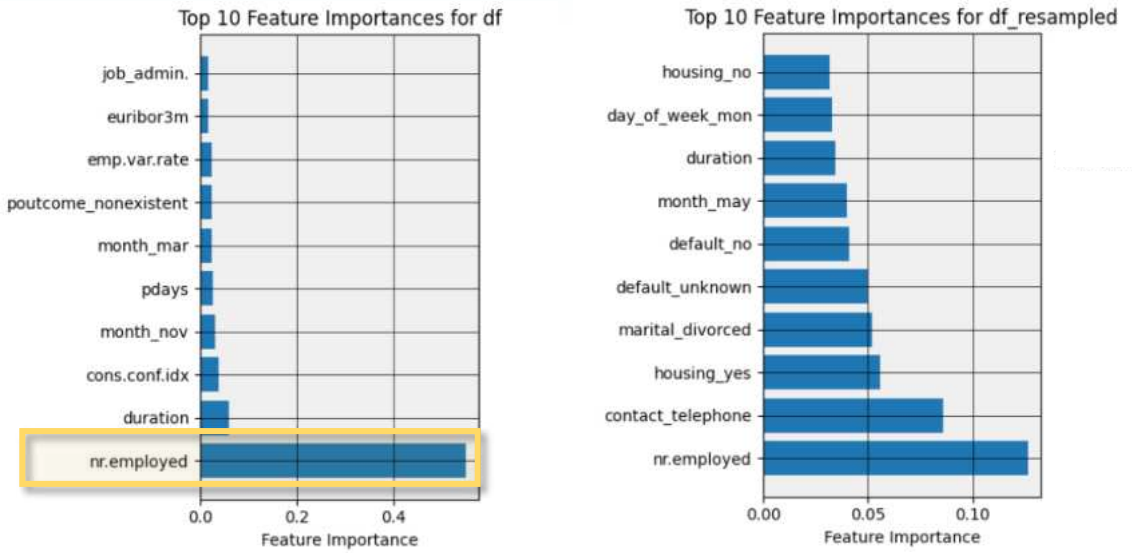


Figure 22 The most influential features of the cleaned and resampled data frames, in the XGBoost classifying models.

Suggested Next Steps

The dataset is considered low-quality. A significant portion of rows contain at least one unknown variable, some fields are unclear (e.g. **pdays:** 999), and the cohort of interest represents 9.4% of the total dataset. It is advised that future campaigns attempt to acquire **complete datasets**, prioritising their **cohort of interest**.

While personal data appears to be bountiful, modelling suggests the key indicators are **financial** and **occupational**-driven, rendering many of these features less imperative. Subsequently, the nature of some occupational-associated answers ranges from **specific** to **broad** (blue-collar vs technician). Given that these features are considered key to the model's success, it's advised to clearly distinguish these professions to help clearly identify target customers.

Alternative features to consider following ethics approval:
"**duration within a particular occupation**", "**salary**", "**residential suburb**", "**household composition**", & **geographical** data.

The following insights are intended to augment the understanding of the known cohort of interest. For associated features of this cohort, see the prior EDA or figures provided in the appendix.

Appendix

Data Dictionary

Variable Name	Description
age	*Age*
job	*Type of job*
marital	*Marital status*
education	*Level of education*
default	*Has credit in default*
balance	*Average yearly balance*
housing	*Has a housing loan*
loan	*Has a personal loan*
contact	*Contact communication type*
day	*Day of contact*
month	*Month of contact*
duration	*Last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y="no"). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known.*
campaign	*Number of contacts performed during this campaign and for this client*
pdays	*Number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)*
previous	*Number of contacts performed before this campaign and for this client*
poutcome	*Outcome of the previous marketing campaign*
emp.var.rate	*employment variation rate - quarterly indicator (numeric)*
cons.price.idx	*consumer price index - monthly indicator (numeric)*
cons.conf.idx	*consumer confidence index - monthly indicator (numeric)*
euribor3m	*euribor 3 month rate - daily indicator (numeric)*
nr.employed	*number employed - quarterly indicator (numeric)*
y	*Did the client subscribe to a Telecom plan? [Feature of interest]*

The dataframe following one-hot-encoding of categorical variables.

Numerical Features (scaled):

1. *age*
2. *duration*
3. *campaign*
4. *pdays*
5. *previous*
6. *emp.var.rate*
7. *cons.price.idx*
8. *cons.conf.idx*
9. *euribor3m*
10. *nr.employed*

Categorical Features and Their One-Hot Encoded Derivatives:

Job Related:

- *job*
- *job_admin.*
- *job_blue-collar*
- *job_entrepreneur*
- *job_housemaid*
- *job_management*
- *job_retired*
- *job_self-employed*
- *job_services*
- *job_student*
- *job_technician*
- *job_unemployed*

Marital Status:

- *marital*
- *marital_divorced*
- *marital_married*
- *marital_single*

Education Level:

- *education*
- *education_basic.4y*
- *education_basic.6y*
- *education_basic.9y*
- *education_high.school*
- *education_illiterate*
- *education_professional.course*
- *education_university.degree*

Default Status:

- *default*
- *default_no*
- *default_yes*

Housing Loan Status:

- *housing*
- *housing_no*
- *housing_yes*

Personal Loan Status:

- *loan*
- *loan_no*
- *loan_yes*

Contact Type:

- *contact*
- *contact_cellular*
- *contact_telephone*

Month of Last Contact:

- *month*
- *month_apr*
- *month_aug*
- *month_dec*
- *month_jul*
- *month_jun*
- *month_mar*
- *month_may*
- *month_nov*
- *month_oct*
- *month_sep*

Day of Week of Last Contact:

- *day_of_week*
- *day_of_week_fri*
- *day_of_week_mon*
- *day_of_week_thu*
- *day_of_week_tue*
- *day_of_week_wed*

Outcome of Previous Marketing Campaign:

- *poutcome*
- *poutcome_failure*
- *poutcome_nonexistent*
- *poutcome_success*

Feature Engineering

Age Binning:

This process categorizes ages into groups such as 'young', 'middle-aged', and 'senior', allowing the model to treat age as a categorical variable rather than numerical, which might capture more nuanced patterns in the age distribution.

Campaign-Related Features:

By calculating the ratio of previous contacts to the number of contacts during the current campaign and creating a binary indicator of whether a client was previously contacted, these features aim to capture the influence of past and current marketing efforts on the outcome.

Economic Indicator Aggregates:

These features provide a mean value of various economic indicators grouped by the age category, potentially highlighting economic conditions that might affect different age groups differently when making a decision.

Month and Day-of-Week Aggregates:

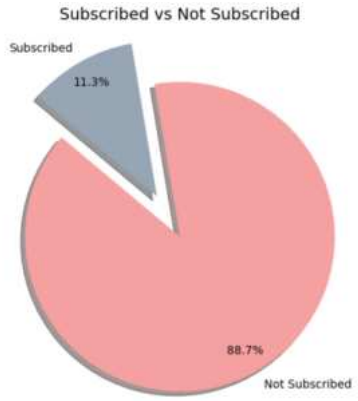
Count variables are created for each month and day of the week, which could help the model understand patterns related to specific times that may be crucial for predicting outcomes.
Contact Type Indicator:

Binary variables are introduced to indicate the type of contact (cellular or non-cellular), which could be an important factor in how clients respond to marketing campaigns.

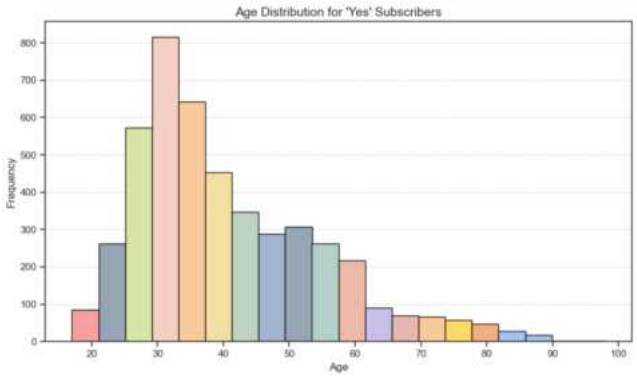
Interaction Features:

By multiplying economic indicators with one another, these features aim to capture the interaction effects between different economic factors and how these combined effects might influence a client's decision.

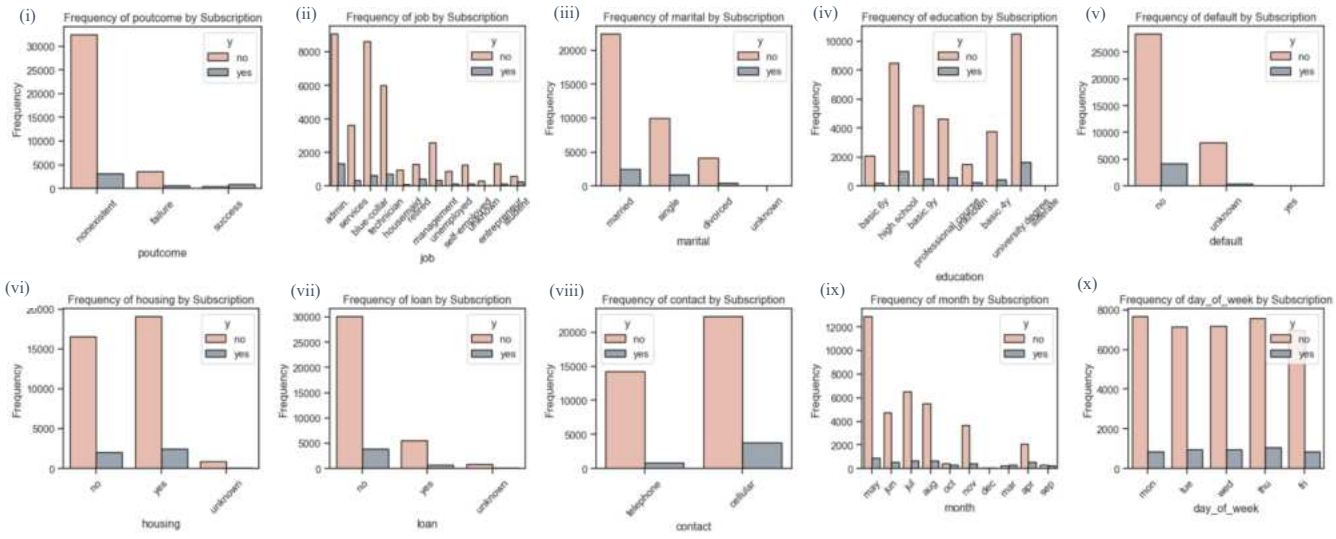
EDA Phase Visualisations



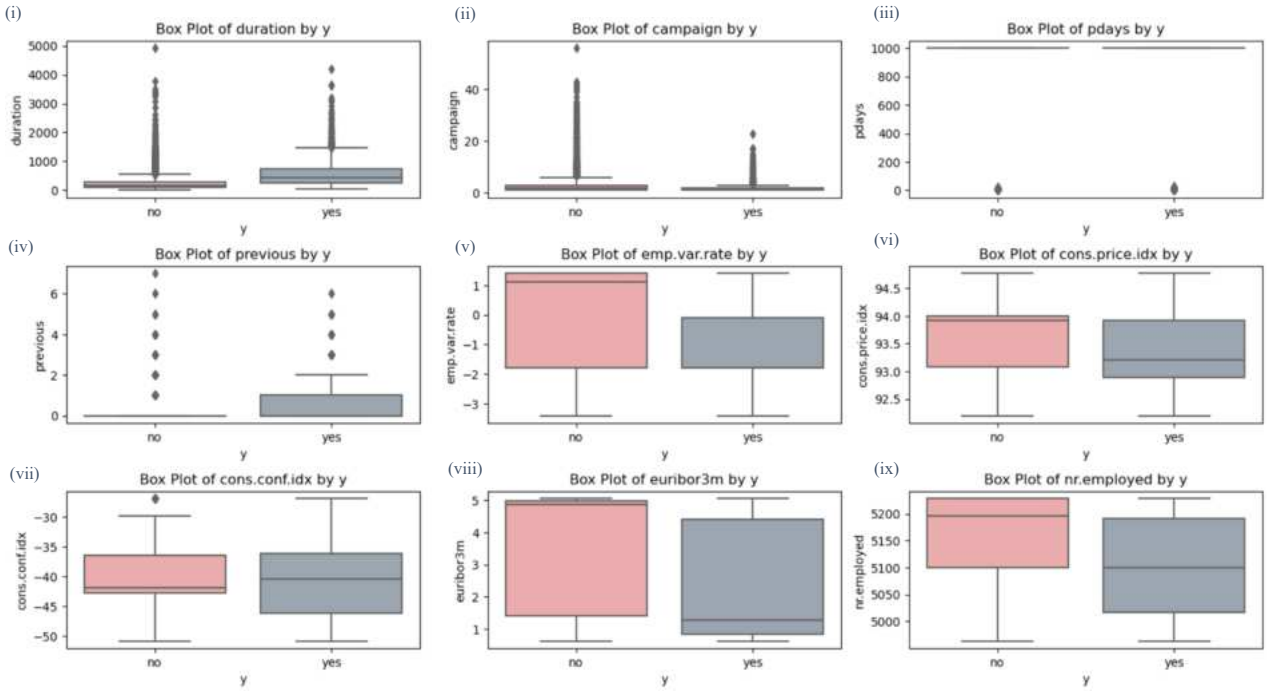
Ap. Figure 1 Pie graph visualisation of the target variable "y".



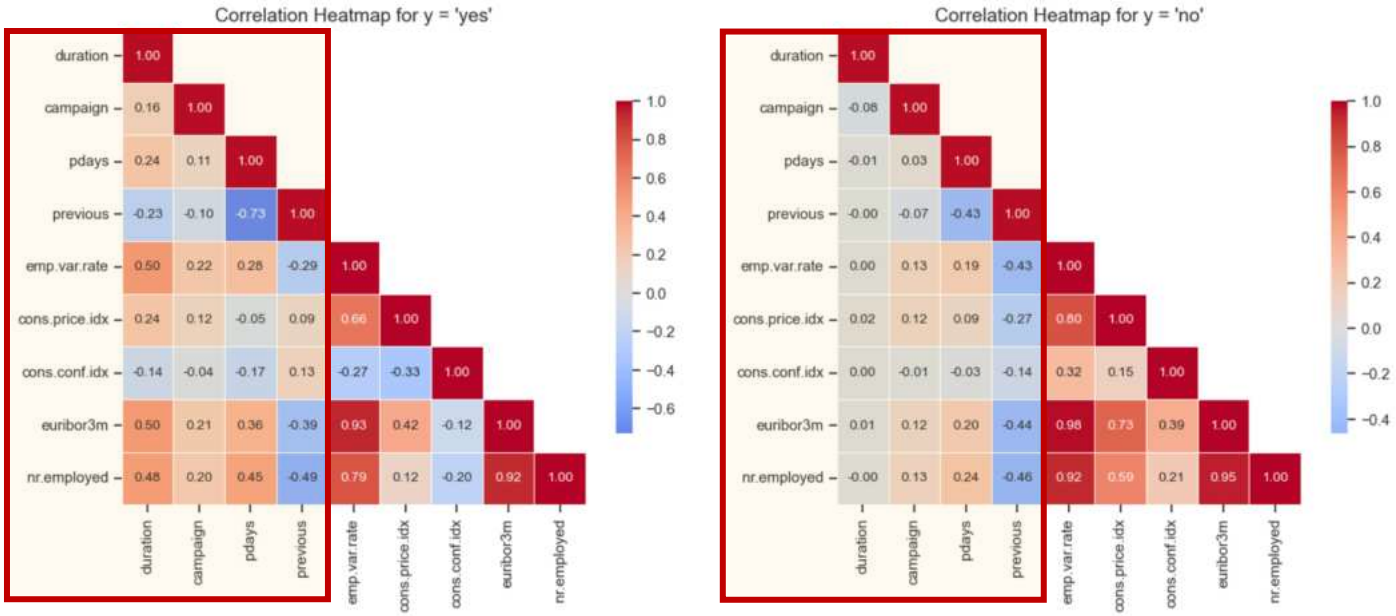
Ap. Figure 2 Bar chart illustrating the age distribution of subscribers.



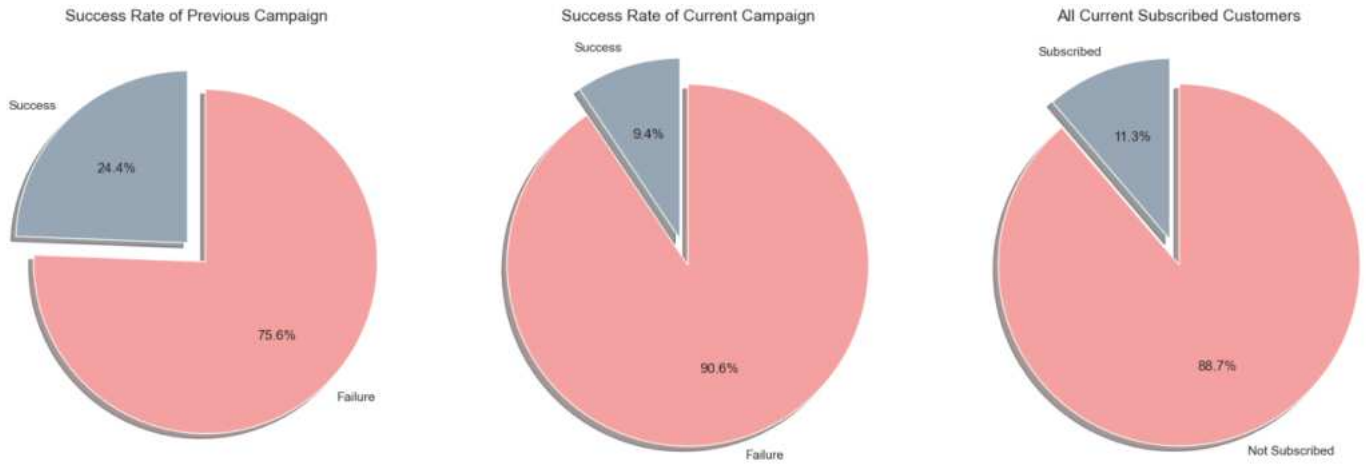
Ap. Figure 3 A series of bar charts illustrating the categorical features of subscribers and non-subscribers.



Ap. Figure 4 A series of box and whisker plots illustrating the numerical features of subscribers and non-subscribers.



Ap. Figure 5 Heatmaps illustrating the correlation between numerical features of subscribers and non-subscribers.



Ap. Figure 6 A series of pie graphs comparing the success rate of the previous campaign, to the current campaign, to all present subscribers. ***Further visuals available in the appendix – see Figures 11 & 12.****.*

	Previous Campaign	Current Campaign	All Current Subscribers
age	29	33	31
job	admin.	admin.	admin.
marital	married	married	married
education	university.degree	university.degree	university.degree
default	no	no	no
housing	yes	yes	yes
loan	no	no	no
contact	cellular	cellular	cellular
month	may	may	may
day_of_week	thu	thu	thu
duration	192	209	301
campaign	1	1	1
pdays	3	999	999
previous	1	0	0
poutcome	success	nonexistent	nonexistent
emp.var.rate	-1.8	-1.8	-1.8
cons.price.idx	92.893	92.893	92.893
cons.conf.idx	-46.2	-46.2	-46.2
euribor3m	0.879	4.962	4.962
nr.employed	4991.6	5099.1	5099.1
y	yes	yes	yes

Ap. Figure 7 A series of data frames comparing the traits of subscribers from the previous campaign, subscribers from the current campaign, and all present subscribers. This figure illustrates the most sensitive to the new campaign is a ***married, early 30s and university educated individual, who has access to housing*, primarily uses cellular and has had at least one contact during the campaign****. This individual is most receptive on a Thursday, at the conclusion of Autumn, and was not exposed to the previous campaign. At this stage, index ranges (bottom 5 variables) don't express immediate significant influences on subscription rates.*

Appendix

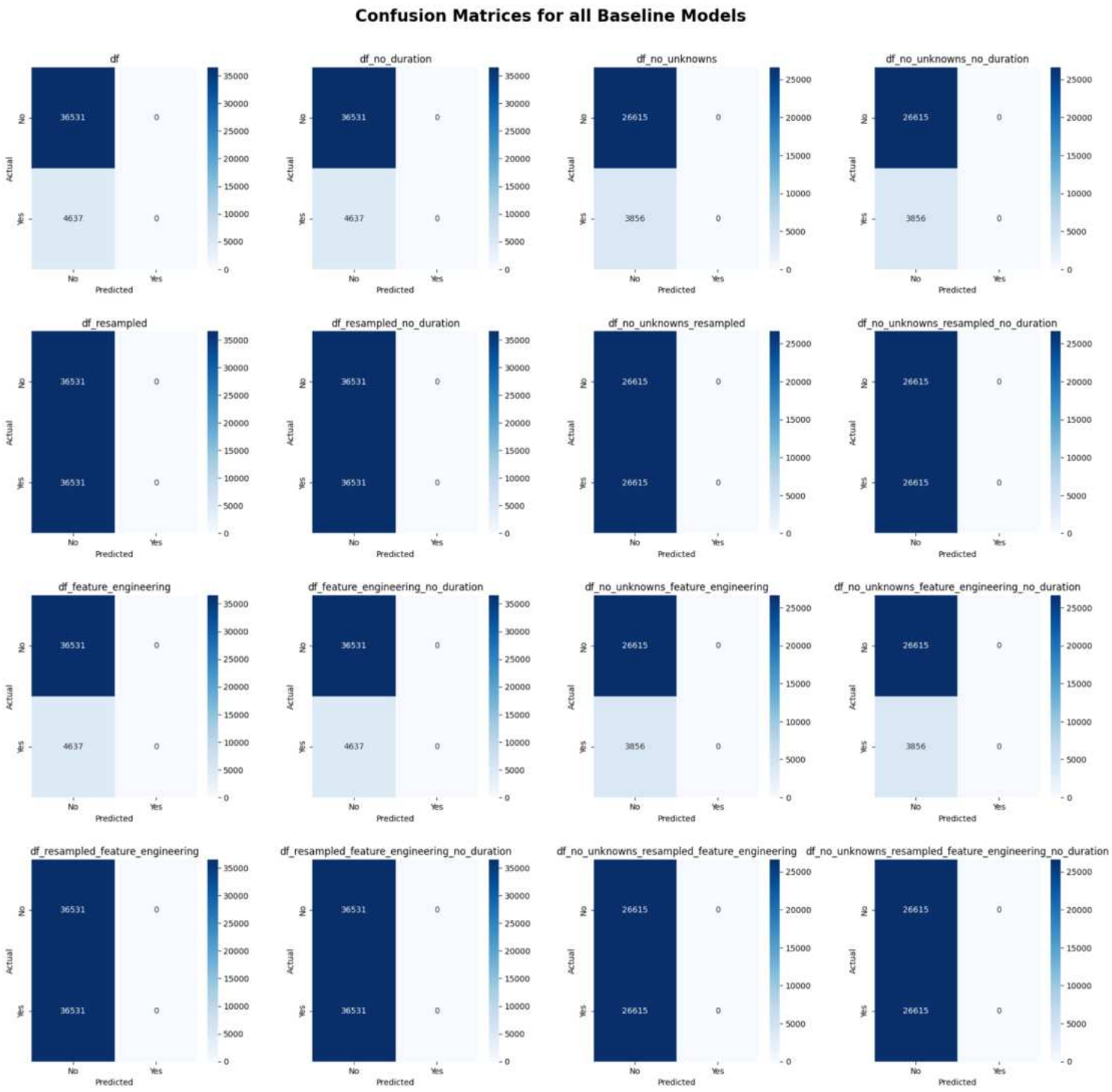
Modelling Phase

DataFrame	Description
df	The original, cleaned dataframe.
df_no_duration	The original, cleaned dataframe without the "duration" feature.
df_no_unknowns	The original, cleaned dataframe (with unknown values omitted).
df_no_unknowns_no_duration	The original, cleaned dataframe with unknown values omitted and without the "duration" feature.
df_resampled	The original, cleaned dataframe (with resampling to account for the unbalanced dataset).
df_resampled_no_duration	The original, cleaned dataframe with resampling to account for the unbalanced dataset and without the "duration" feature.
df_no_unknowns_resampled	The original, cleaned dataframe (with unknown values omitted and resampling to account for the unbalanced dataset).
df_no_unknowns_resampled_no_duration	The original, cleaned dataframe with unknown values omitted, resampling for the unbalanced dataset, and without the "duration" feature.
df_feature_engineering	The original, cleaned dataframe (with feature engineering applied).
df_feature_engineering_no_duration	The original, cleaned dataframe with feature engineering applied and without the "duration" feature.
df_no_unknowns_feature_engineering	The original, cleaned dataframe (with unknown values omitted and feature engineering applied).
df_no_unknowns_feature_engineering_no_duration	The original, cleaned dataframe with unknown values omitted, feature engineering applied, and without the "duration" feature.
df_resampled_feature_engineering	The original, cleaned dataframe (with resampling to account for the unbalanced dataset and feature engineering applied).
df_resampled_feature_engineering_no_duration	The original, cleaned dataframe with resampling for the unbalanced dataset, feature engineering applied, and without the "duration" feature.
df_no_unknowns_resampled_feature_engineering	The original, cleaned dataframe (with unknown values omitted, resampling to account for the unbalanced dataset and feature engineering applied).
df_no_unknowns_resampled_feature_engineering_no_duration	The original, cleaned dataframe with unknown values omitted, resampling for the unbalanced dataset, feature engineering applied, and without the "duration" feature.

Ap. Figure 8 The complete set of modelled dataframes.

DataFrame	Baseline Accuracy
df	88.74%
df_no_duration	88.74%
df_no_unknowns	87.35%
df_no_unknowns_no_duration	87.35%
df_resampled	50.00%
df_resampled_no_duration	50.00%
df_no_unknowns_resampled	50.00%
df_no_unknowns_resampled_no_duration	50.00%
df_feature_engineering	88.74%
df_feature_engineering_no_duration	88.74%
df_no_unknowns_feature_engineering	87.35%
df_no_unknowns_feature_engineering_no_duration	87.35%
df_resampled_feature_engineering	50.00%
df_resampled_feature_engineering_no_duration	50.00%
df_no_unknowns_resampled_feature_engineering	50.00%
df_no_unknowns_resampled_feature_engineering_no_duration	50.00%

Ap. Figure 9 Baseline Model Accuracies for each data frame.



Ap. Figure 10 Confusion matrix visualisations per corresponding baseline model.

Training Metrics Table

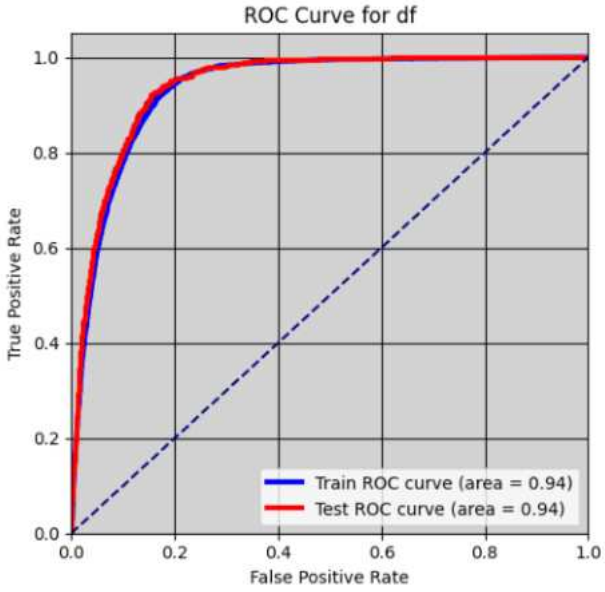
	DataFrame	Accuracy	AUC	Error Rate	Sensitivity	Specificity	Precision	Recall	Log Loss
0	df	0.910852	0.935065	0.089148	0.421661	0.973163	0.666808	0.421661	0.208759
1	df_no_duration	0.899405	0.795683	0.100595	0.228971	0.984801	0.657407	0.228971	0.276057
2	df_no_unknowns	0.899286	0.928467	0.100714	0.432536	0.968175	0.667323	0.432536	0.232021
3	df_no_unknowns_no_duration	0.887717	0.802449	0.112283	0.245614	0.982487	0.674256	0.245614	0.297959
4	df_resampled	0.949837	0.991735	0.050163	0.926773	0.972900	0.971588	0.926773	0.117039
5	df_resampled_no_duration	0.943917	0.974028	0.056083	0.902135	0.985697	0.984393	0.902135	0.155756
6	df_no_unknowns_resampled	0.942748	0.989672	0.057252	0.917109	0.968390	0.966685	0.917109	0.131971
7	df_no_unknowns_resampled_no_duration	0.936408	0.971309	0.063592	0.889823	0.982998	0.981252	0.889823	0.169754
8	df_feature_engineering	0.910882	0.935931	0.089118	0.424080	0.972889	0.665823	0.424080	0.207887
9	df_feature_engineering_no_duration	0.899223	0.795944	0.100777	0.227896	0.984733	0.655332	0.227896	0.275621
10	df_no_unknowns_feature_engineering	0.899163	0.929178	0.100837	0.433174	0.967939	0.666013	0.433174	0.231434
11	df_no_unknowns_feature_engineering_no_duration	0.887266	0.802418	0.112734	0.245295	0.982016	0.668115	0.245295	0.297743
12	df_resampled_feature_engineering	0.949631	0.991743	0.050369	0.926875	0.972387	0.971069	0.926875	0.117333
13	df_resampled_feature_engineering_no_duration	0.943387	0.973759	0.056613	0.902341	0.984431	0.983038	0.902341	0.156442
14	df_no_unknowns_resampled_feature_engineering	0.942537	0.989666	0.057463	0.917297	0.967780	0.966070	0.917297	0.132352
15	df_no_unknowns_resampled_feature_engineering_n...	0.935962	0.971047	0.064038	0.890340	0.981588	0.979742	0.890340	0.170658

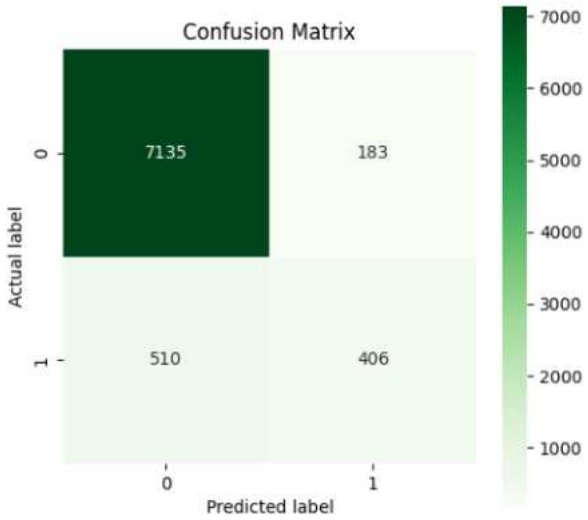
Testing Metrics Table

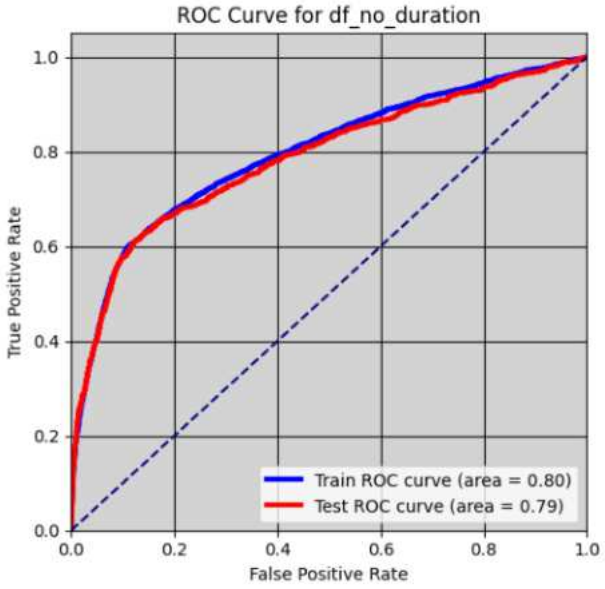
	DataFrame	Accuracy	AUC	Error Rate	Sensitivity	Specificity	Precision	Recall	Log Loss
0	df	0.915837	0.939617	0.084163	0.443231	0.974993	0.689304	0.443231	0.202501
1	df_no_duration	0.903206	0.786583	0.096794	0.241266	0.986062	0.684211	0.241266	0.275317
2	df_no_unknowns	0.908942	0.935694	0.091058	0.443828	0.971344	0.675105	0.443828	0.213096
3	df_no_unknowns_no_duration	0.895980	0.794649	0.104020	0.241331	0.983811	0.666667	0.241331	0.286768
4	df_resampled	0.949702	0.991663	0.050298	0.926098	0.973310	0.971991	0.926098	0.120398
5	df_resampled_no_duration	0.941969	0.972896	0.058031	0.899548	0.984396	0.982952	0.899548	0.158187
6	df_no_unknowns_resampled	0.942608	0.989851	0.057392	0.915633	0.969572	0.967825	0.915633	0.129562
7	df_no_unknowns_resampled_no_duration	0.934999	0.970499	0.065001	0.887824	0.982156	0.980290	0.887824	0.172051
8	df_feature_engineering	0.914744	0.939931	0.085256	0.443231	0.973763	0.678930	0.443231	0.202720
9	df_feature_engineering_no_duration	0.902356	0.786366	0.097644	0.237991	0.985515	0.672840	0.237991	0.275631
10	df_no_unknowns_feature_engineering	0.908450	0.937159	0.091550	0.443828	0.970785	0.670860	0.443828	0.211753
11	df_no_unknowns_feature_engineering_no_duration	0.895160	0.795888	0.104840	0.239945	0.983067	0.655303	0.239945	0.286166
12	df_resampled_feature_engineering	0.949360	0.991649	0.050640	0.925688	0.973036	0.971699	0.925688	0.120979
13	df_resampled_feature_engineering_no_duration	0.941422	0.972506	0.058578	0.899138	0.983712	0.982210	0.899138	0.159006
14	df_no_unknowns_resampled_feature_engineering	0.943077	0.989911	0.056923	0.916009	0.970135	0.968415	0.916009	0.129574
15	df_no_unknowns_resampled_feature_engineering_n...	0.933872	0.970248	0.066128	0.888388	0.979339	0.977263	0.888388	0.172948

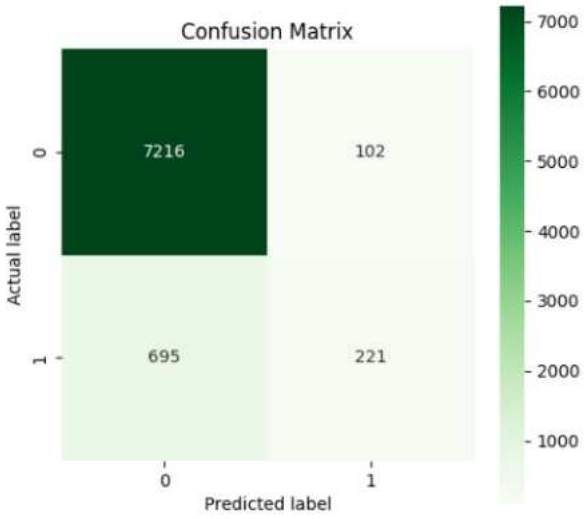
Ap. Figure 11 Logistic Regression Performance metrics per data frame.

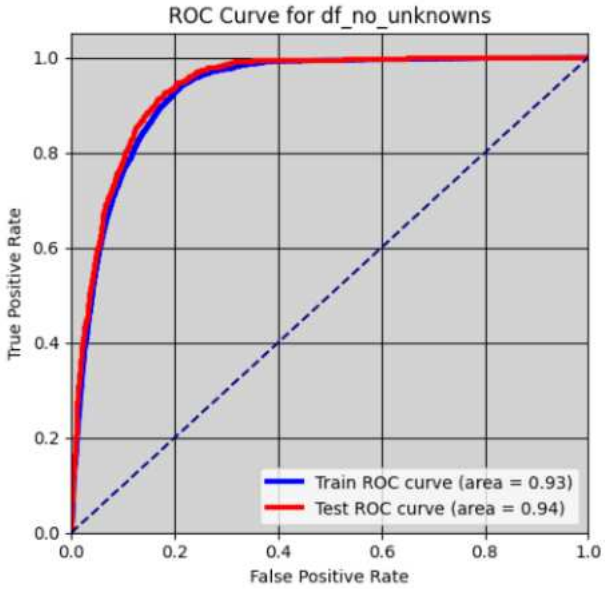
Logistic Regression Iterations performed per data frame.

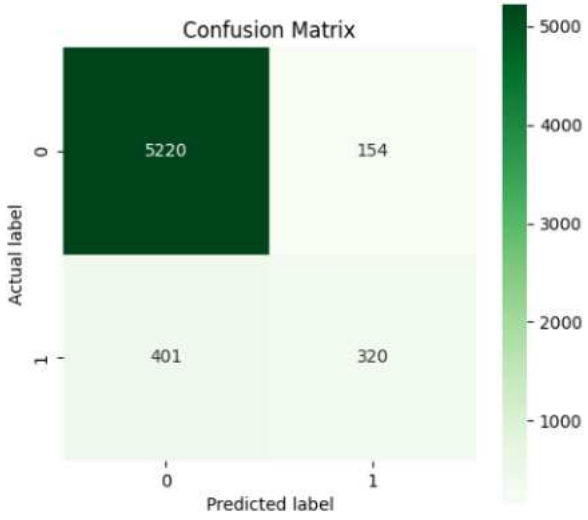


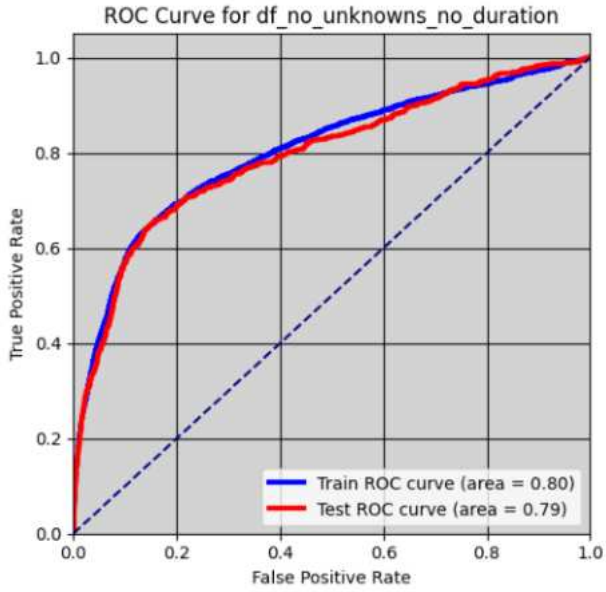


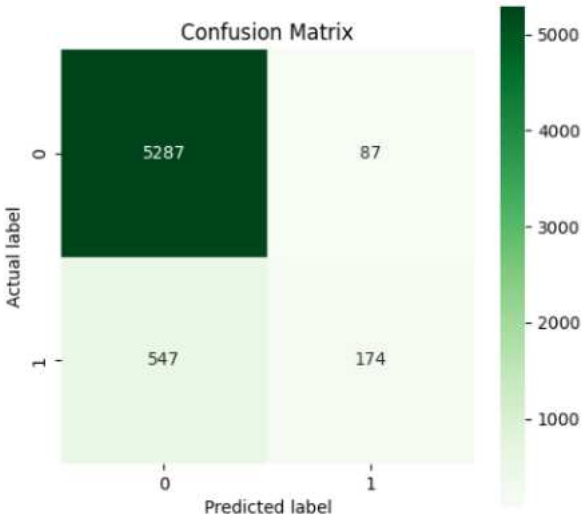


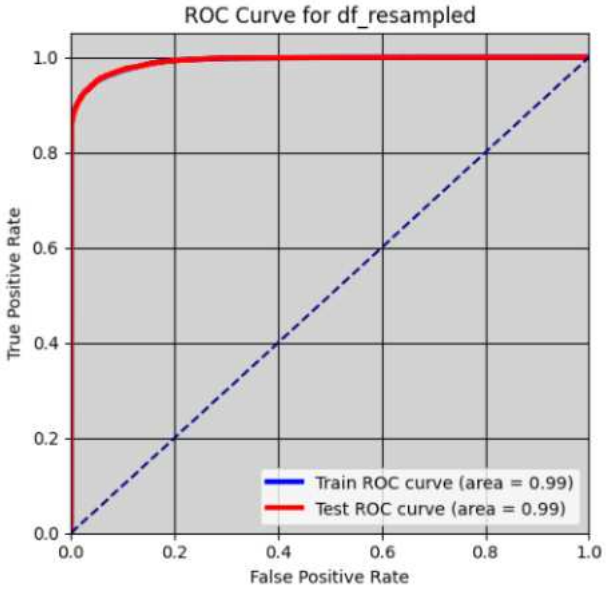


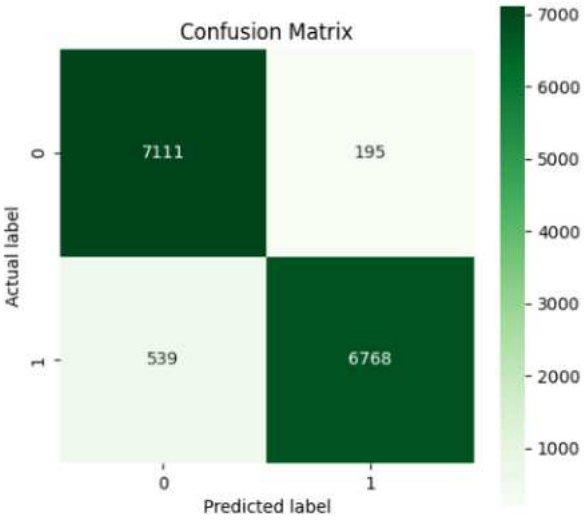


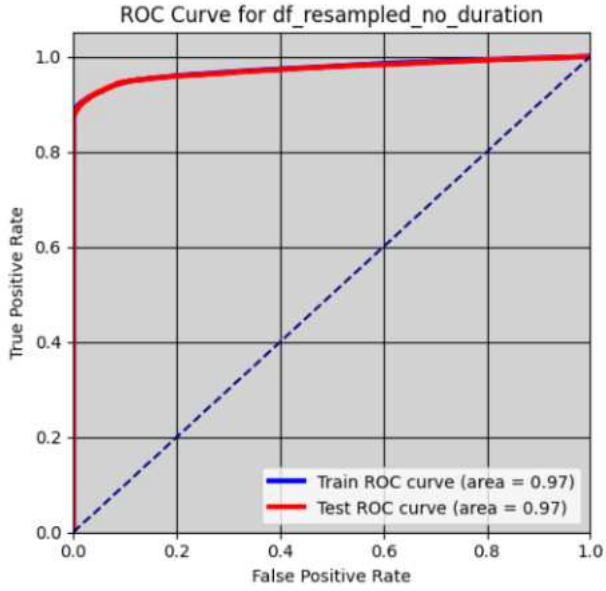


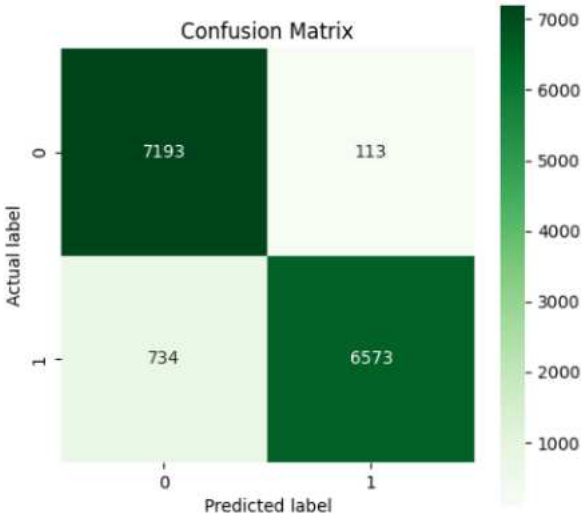

ROC Curve for df_no_unknowns_no_duration
True Positive Rate
False Positive Rate
Train ROC curve (area = 0.80)
Test ROC curve (area = 0.79)

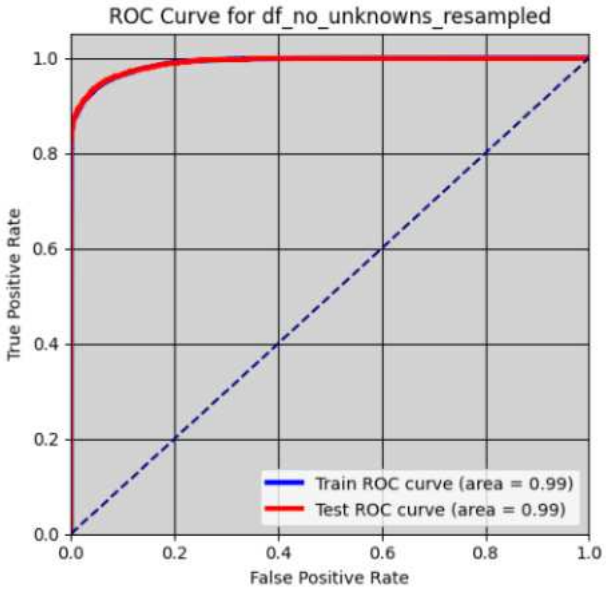

Confusion Matrix
Actual label
Predicted label
5287
87
547
174

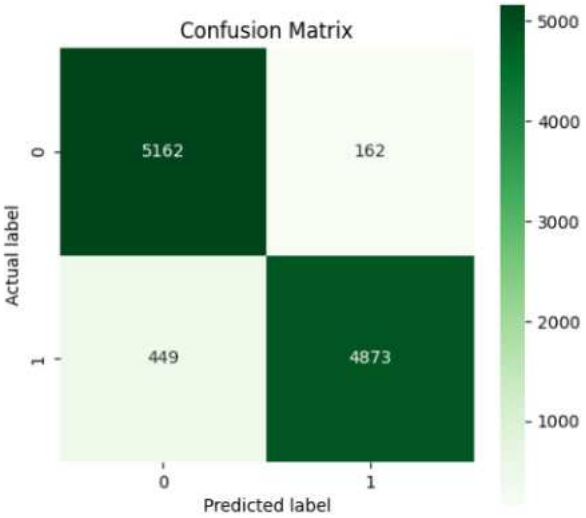

ROC Curve for df_resampled
True Positive Rate
False Positive Rate
Train ROC curve (area = 0.99)
Test ROC curve (area = 0.99)

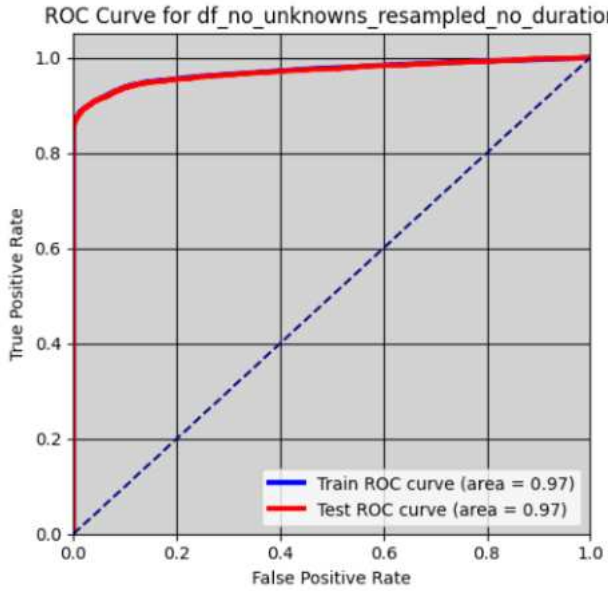

Confusion Matrix
Actual label
Predicted label
7111
195
539
6768

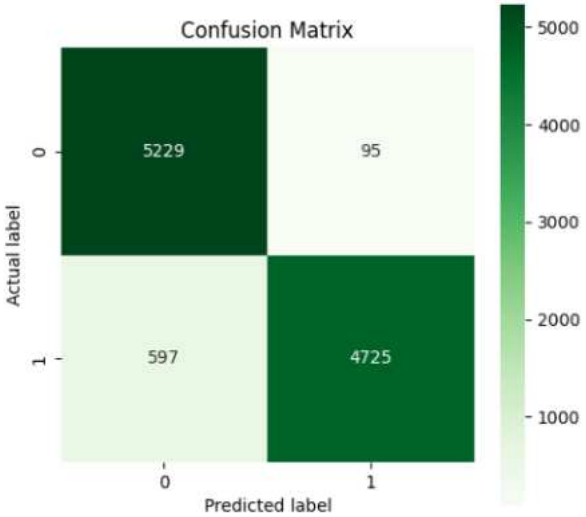

ROC Curve for df_resampled_no_duration
True Positive Rate
False Positive Rate
Train ROC curve (area = 0.97)
Test ROC curve (area = 0.97)

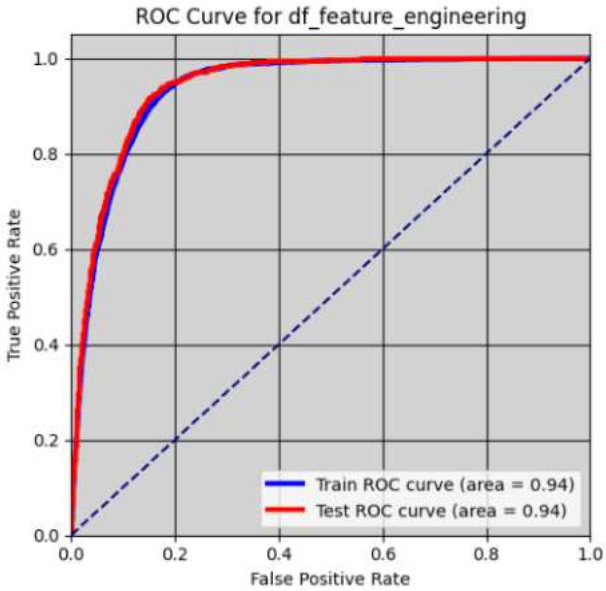

Confusion Matrix
Actual label
Predicted label
7193
113
734
6573

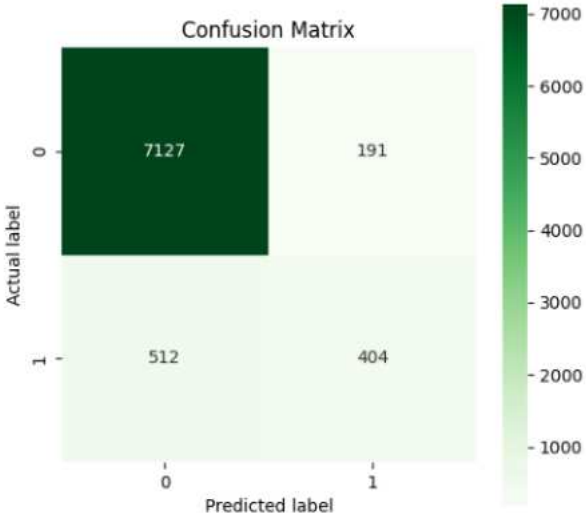

ROC Curve for df_no_unknowns_resampled
True Positive Rate
False Positive Rate
Train ROC curve (area = 0.99)
Test ROC curve (area = 0.99)

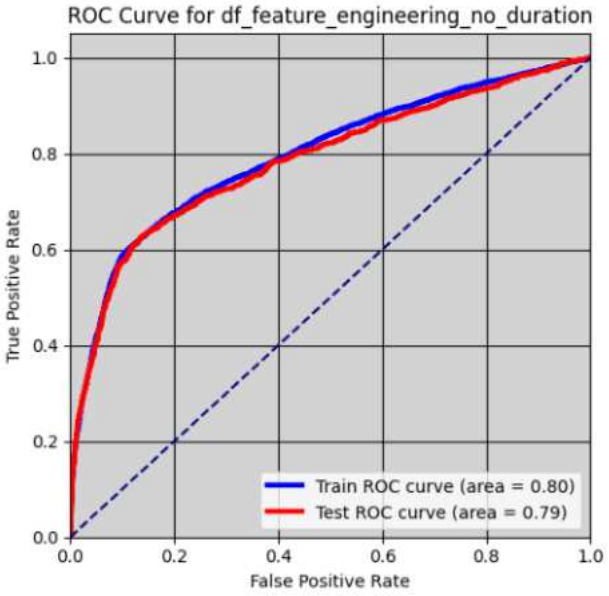

Confusion Matrix
Actual label
Predicted label
5162
162
449
4873

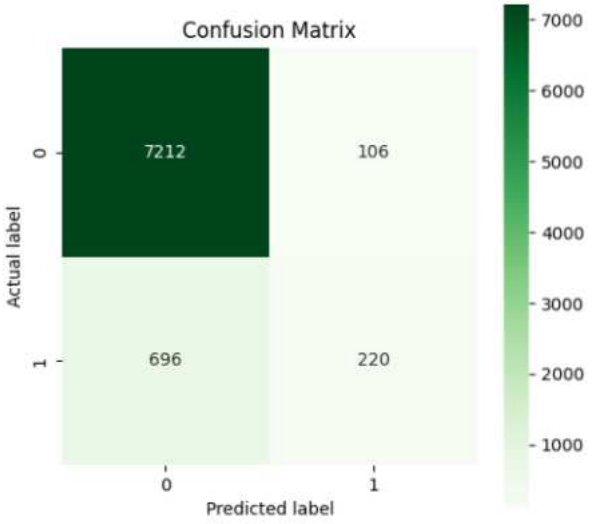

ROC Curve for df_no_unknowns_resampled_no_duration
True Positive Rate
False Positive Rate
Train ROC curve (area = 0.97)
Test ROC curve (area = 0.97)

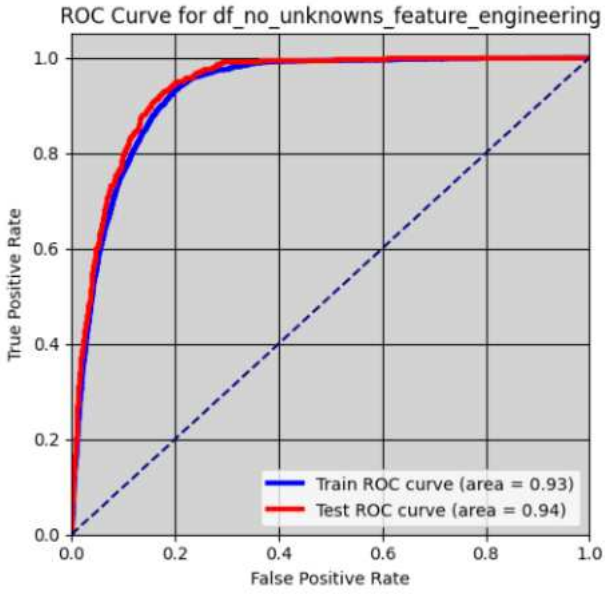

Confusion Matrix
Actual label
Predicted label
5229
95
597
4725

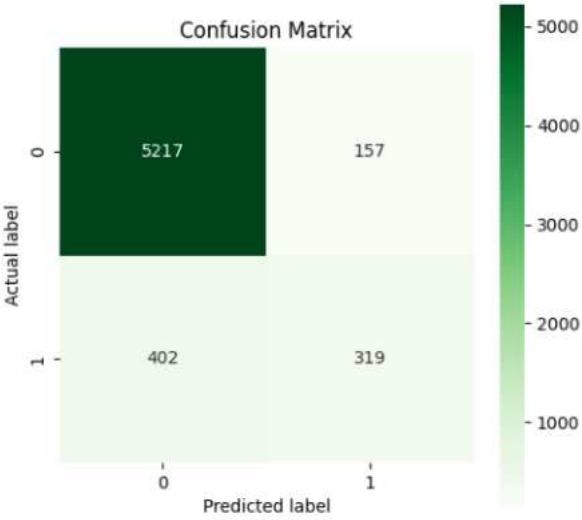

ROC Curve for df_feature_engineering
True Positive Rate
False Positive Rate
Train ROC curve (area = 0.94)
Test ROC curve (area = 0.94)

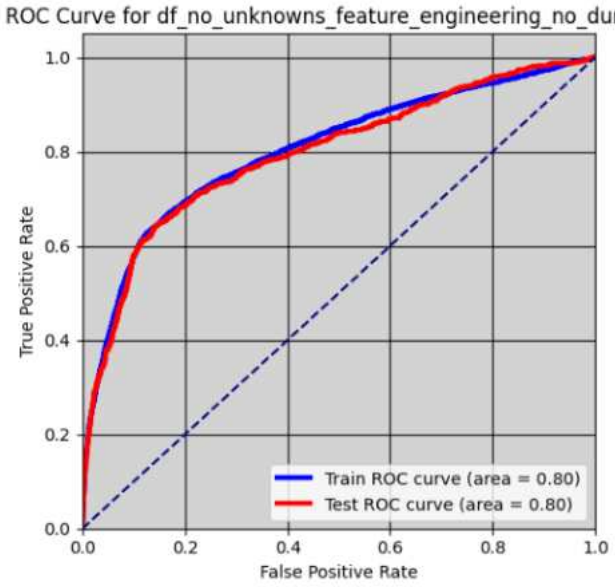

Confusion Matrix
Actual label
Predicted label
7127
191
512
404

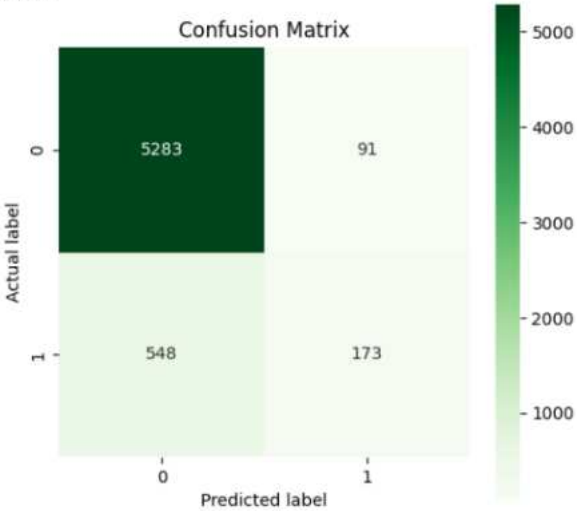

ROC Curve for df_feature_engineering_no_duration
True Positive Rate
False Positive Rate
Train ROC curve (area = 0.80)
Test ROC curve (area = 0.79)

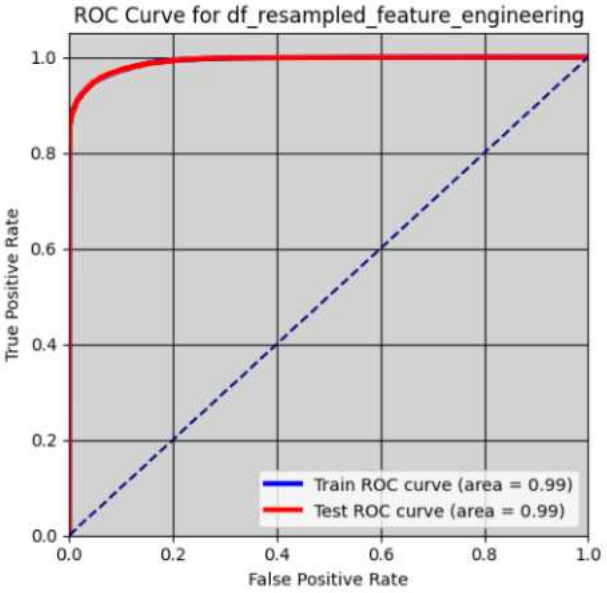

Confusion Matrix
7212
106
696
220
Actual label
Predicted label

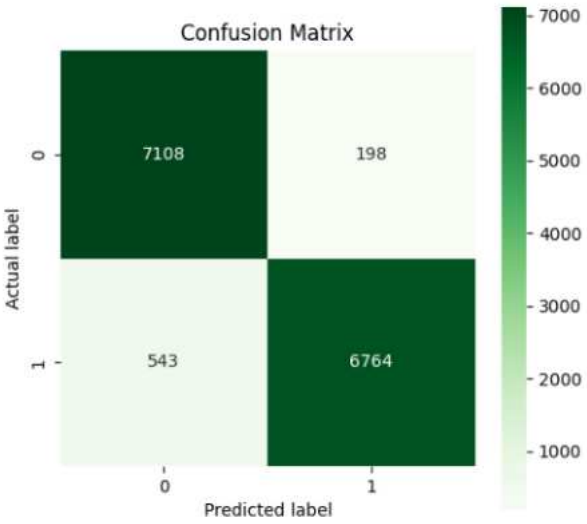

ROC Curve for df_no_unknowns_feature_engineering
True Positive Rate
False Positive Rate
Train ROC curve (area = 0.93)
Test ROC curve (area = 0.94)

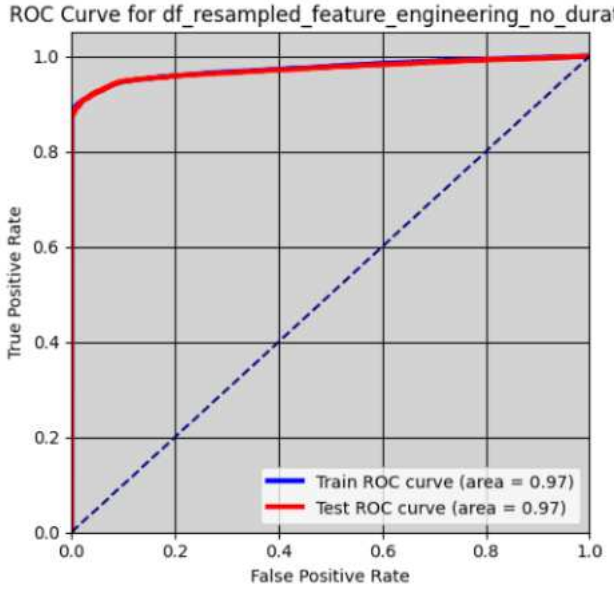

Confusion Matrix
5217
157
402
319
Actual label
Predicted label

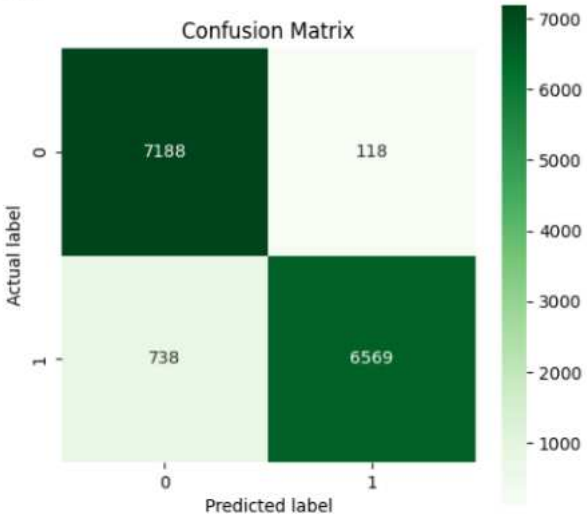

ROC Curve for df_no_unknowns_feature_engineering_no_duration
True Positive Rate
False Positive Rate
Train ROC curve (area = 0.80)
Test ROC curve (area = 0.80)

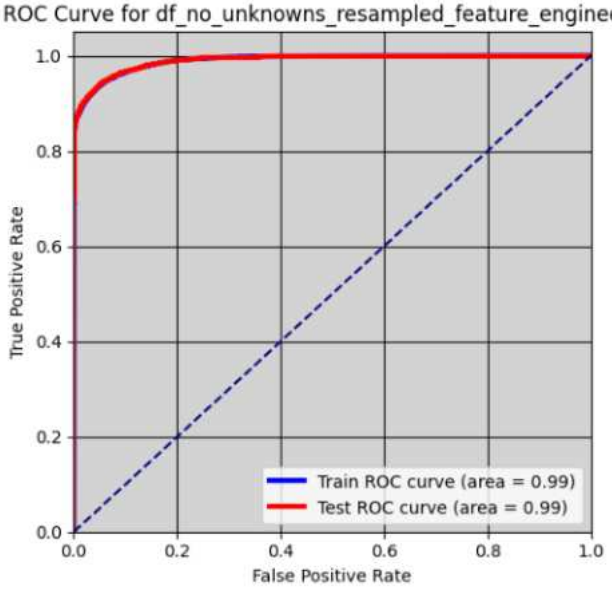

Confusion Matrix
5283
91
548
173
Actual label
Predicted label

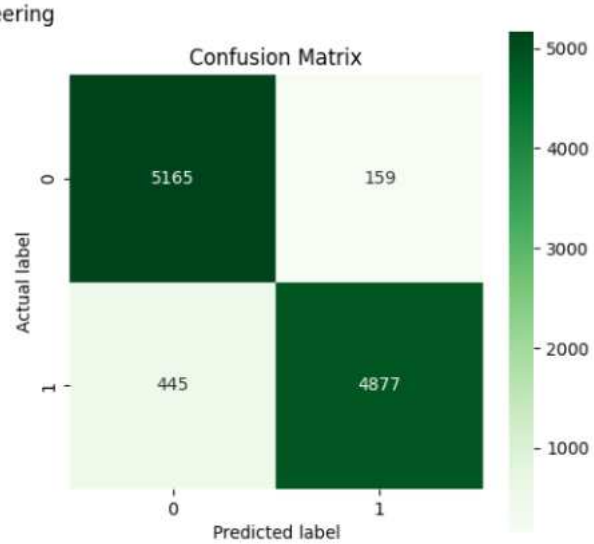

ROC Curve for df_resampled_feature_engineering
True Positive Rate
False Positive Rate
Train ROC curve (area = 0.99)
Test ROC curve (area = 0.99)

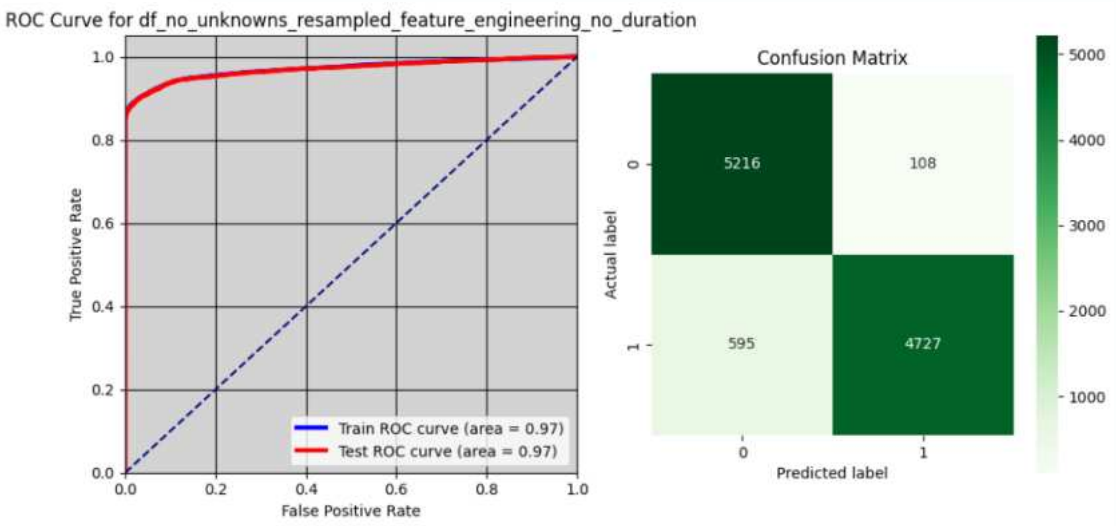

Confusion Matrix
7108
198
543
6764
Actual label
Predicted label


ROC Curve for df_resampled_feature_engineering_no_duration
True Positive Rate
False Positive Rate
Train ROC curve (area = 0.97)
Test ROC curve (area = 0.97)

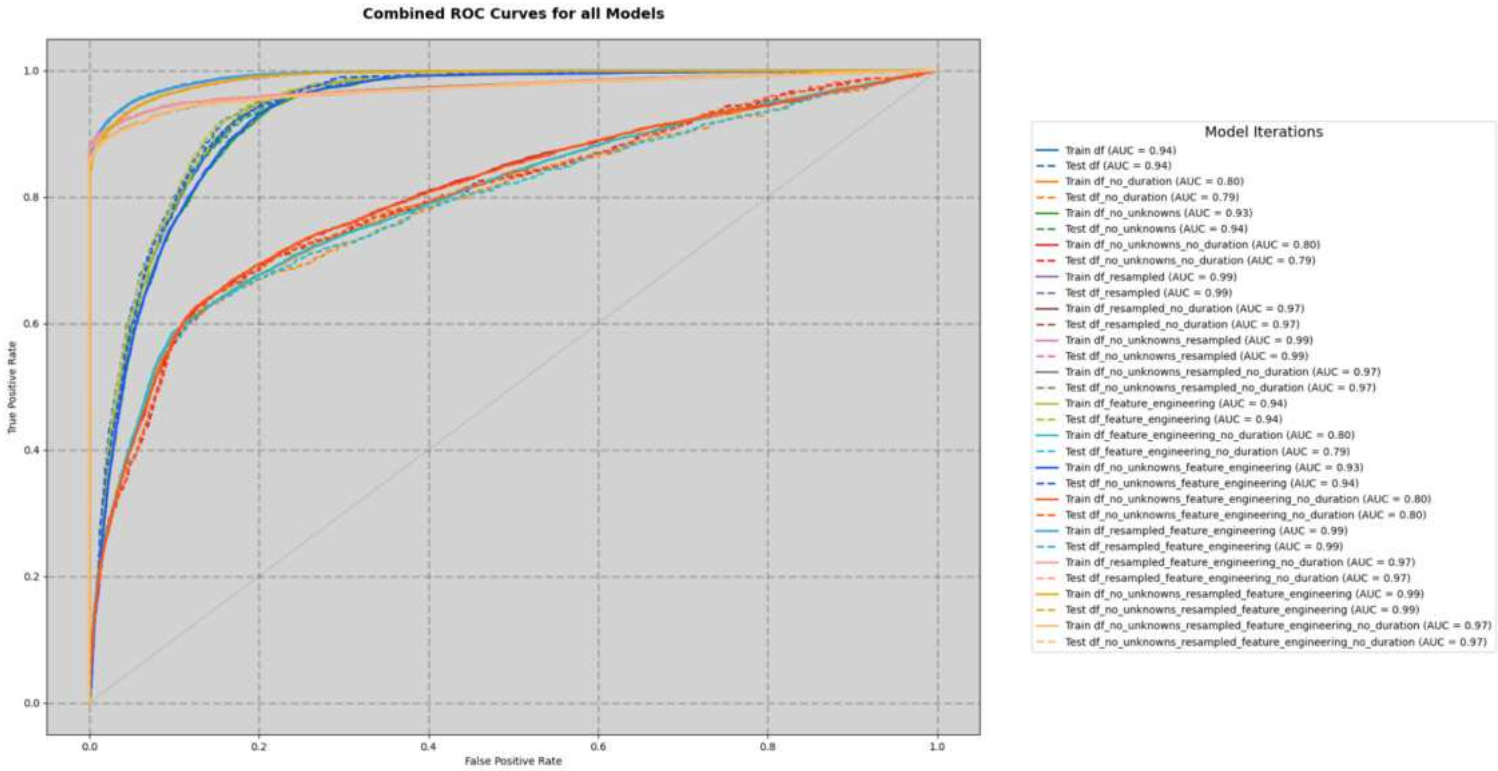

Confusion Matrix
7188
118
738
6569
Actual label
Predicted label

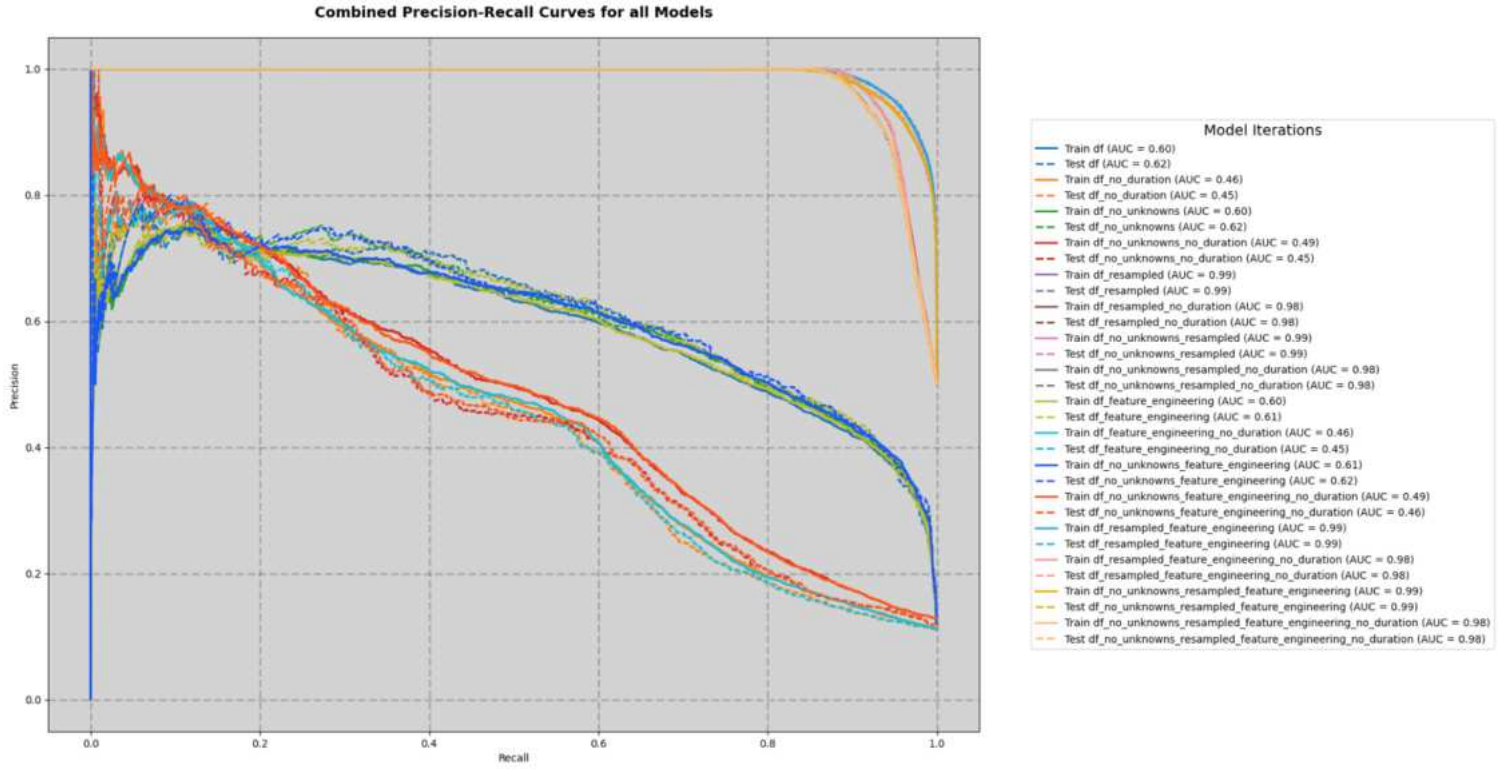

ROC Curve for df_no_unknowns_resampled_feature_engineering
True Positive Rate
False Positive Rate
Train ROC curve (area = 0.99)
Test ROC curve (area = 0.99)


Confusion Matrix
5165
159
445
4877
Actual label
Predicted label

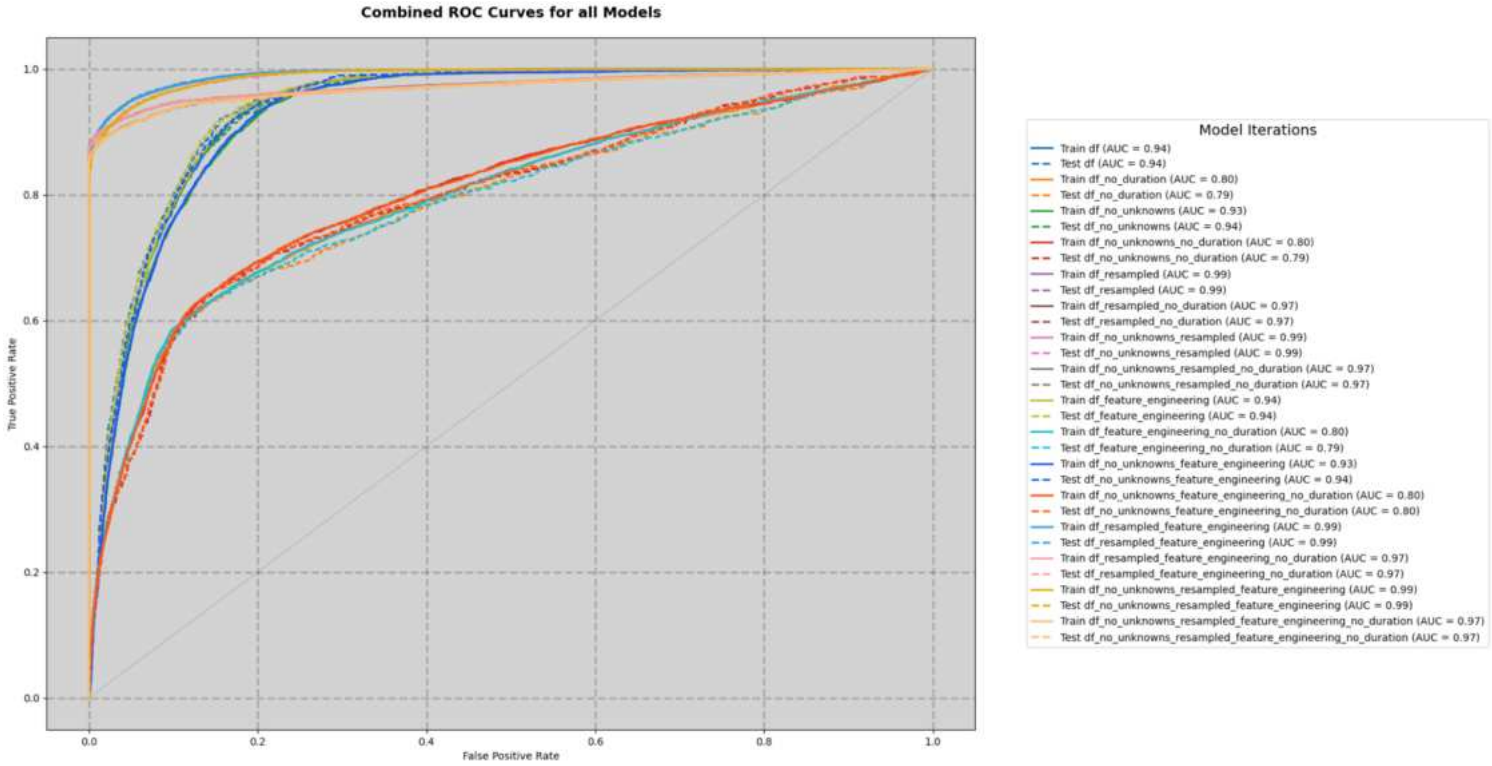


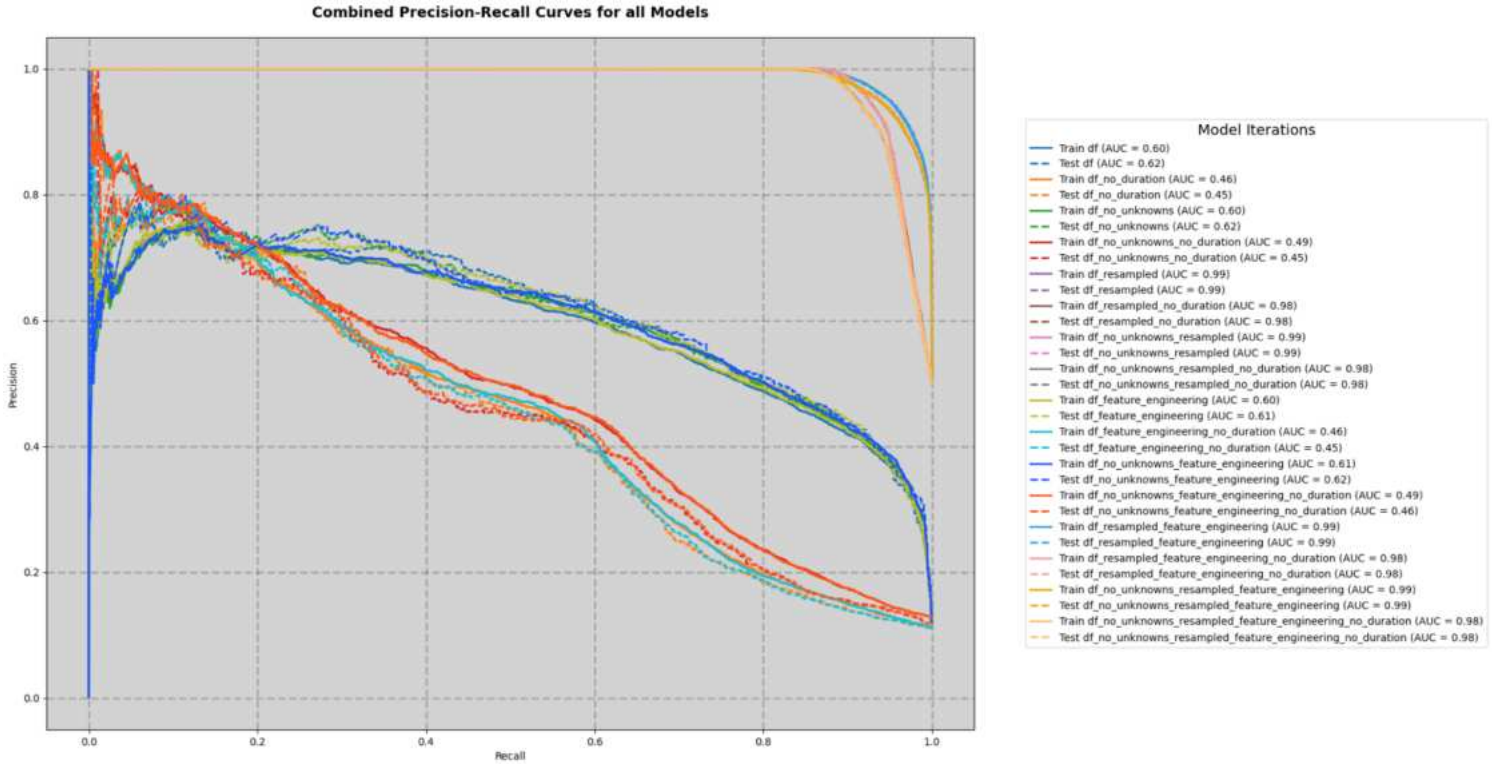
Ap. Figure 12 (set) Logistic Regression Model iterations performed per data frame, with corresponding Confusion matrix.



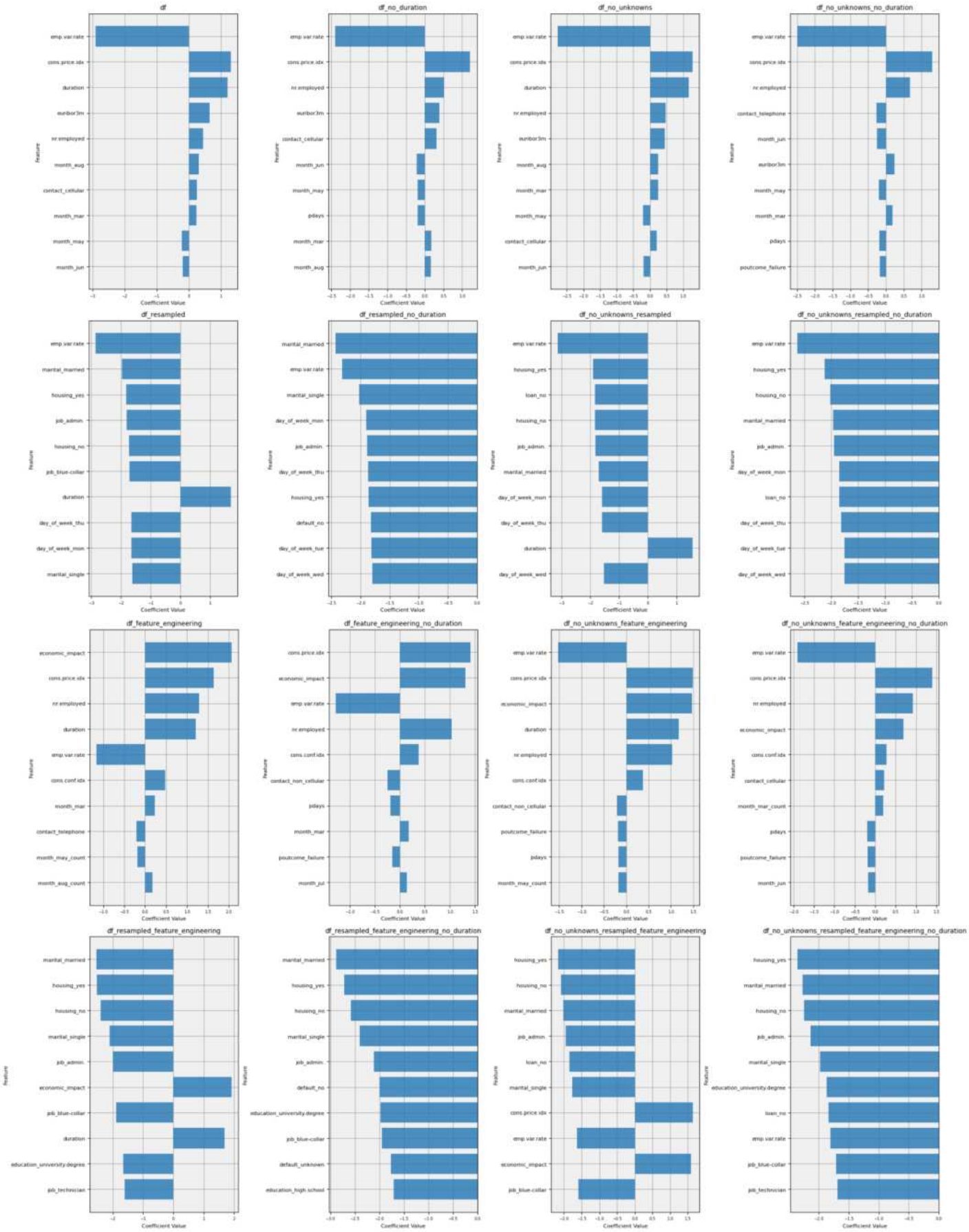


Ap. Figure 13 (set) Training and Testing Standard Logistic Regression Model iterations, ROC curves and Precision-Recall curves.





Ap. Figure 14 (set) Training and Testing Tuned Logistic Regression Model iterations, ROC curves and Precision-Recall curves – (params: penalty: L1, C: 1, Solver: liblinear.



Ap. Figure 15 (set) Coefficient Values of each tuned Logistic Regression model iteration.

Logistic Regression statistically significant features, per data frame.

Summary for df (only statistically significant coefficients at 99%):

	Feature	Coefficient	Lower_CI_99	Upper_CI_99	Statistical_Significance_99
0	age	0.007172	0.000984	0.013360	True
1	duration	0.004645	0.004456	0.004834	True
2	campaign	0.045743	0.022947	0.068538	True
3	pdays	-0.001523	-0.002086	-0.000960	True
4	cons.conf.idx	0.048218	0.027496	0.068940	True

Summary for df_feature_engineering (only statistically significant coefficients at 99%):

	Feature	Coefficient	Lower_CI_99	Upper_CI_99	Statistical_Significance_99
0	age	0.017352	0.009878	0.024826	True
1	duration	0.004487	0.004303	0.004670	True
2	pdays	-0.001762	-0.002320	-0.001204	True

Summary for df_feature_engineering_no_duration (only statistically significant coefficients at 99%):

	Feature	Coefficient	Lower_CI_99	Upper_CI_99	Statistical_Significance_99
0	pdays	-0.001712	-0.002235	-0.001189	True

Summary for df_no_duration (only statistically significant coefficients at 99%):

	Feature	Coefficient	Lower_CI_99	Upper_CI_99	Statistical_Significance_99
0	campaign	-0.064125	-0.090612	-0.037638	True
1	pdays	-0.001853	-0.002397	-0.001308	True
2	cons.conf.idx	0.035367	0.016444	0.054291	True

Summary for df_no_unknowns (only statistically significant coefficients at 99%):

	Feature	Coefficient	Lower_CI_99	Upper_CI_99	Statistical_Significance_99
0	duration	0.004433	0.004222	0.004644	True
1	campaign	0.039334	0.013181	0.065486	True
2	pdays	-0.001518	-0.002120	-0.000917	True
3	cons.conf.idx	0.033281	0.011058	0.055504	True

Summary for df_no_unknowns_feature_engineering (only statistically significant coefficients at 99%):

	Feature	Coefficient	Lower_CI_99	Upper_CI_99	Statistical_Significance_99
0	duration	0.004333	0.004125	0.004540	True
1	pdays	-0.001648	-0.002243	-0.001052	True

Summary for df_no_unknowns_feature_engineering_no_duration (only statistically significant coefficients at 99%):

	Feature	Coefficient	Lower_CI_99	Upper_CI_99	Statistical_Significance_99
0	pdays	-0.001595	-0.002159	-0.001031	True

Summary for df_no_unknowns_no_duration (only statistically significant coefficients at 99%):

	Feature	Coefficient	Lower_CI_99	Upper_CI_99	Statistical_Significance_99
0	campaign	0.025861	0.005778	0.045944	True
1	pdays	-0.001518	-0.002085	-0.000952	True
2	cons.conf.idx	0.022471	0.002735	0.042206	True

Summary for df_no_unknowns_resampled (only statistically significant coefficients at 99%):

	Feature	Coefficient	Lower_CI_99	Upper_CI_99	Statistical_Significance_99
0	duration	0.005474	0.005274	0.005674	True
1	campaign	-0.720875	-0.760060	-0.681690	True
2	pdays	-0.002284	-0.002811	-0.001757	True
3	previous	-0.742073	-0.897416	-0.586731	True
4	emp.var.rate	-0.367354	-0.598958	-0.135750	True
5	cons.price.idx	0.546496	0.154916	0.938077	True
6	cons.conf.idx	0.086077	0.071032	0.101122	True
7	nr.employed	-0.008281	-0.013273	-0.003289	True
8	job_admin.	-0.574897	-0.730548	-0.419246	True
9	job_blue-collar	-0.701947	-0.908470	-0.495424	True
10	job_management	-0.250563	-0.481427	-0.019699	True
11	job_services	-0.367346	-0.600825	-0.133867	True
12	job_technician	-0.415122	-0.602820	-0.227424	True
13	marital_divorced	-0.369176	-0.572216	-0.166135	True
14	marital_married	-0.889530	-1.031745	-0.747316	True
15	marital_single	-0.485573	-0.638270	-0.332877	True
16	education_basic.4y	-0.250433	-0.491167	-0.009700	True
17	education_basic.9y	-0.523365	-0.732344	-0.314386	True
18	education_high.school	-0.772516	-0.941813	-0.603218	True
19	education_professional.course	-0.388178	-0.588843	-0.187512	True
20	education_university.degree	-0.439698	-0.587403	-0.291992	True
21	housing_no	-0.936338	-1.082342	-0.790334	True
22	housing_yes	-0.882849	-1.022867	-0.742832	True
23	loan_no	-0.349378	-0.512596	-0.186159	True
24	loan_yes	-0.516363	-0.726355	-0.306370	True
25	contact_telephone	-0.591739	-0.815559	-0.367920	True
26	month_may	-1.370831	-1.540132	-1.201529	True
27	month_nov	-0.283344	-0.501244	-0.065445	True
28	day_of_week_fri	-0.599340	-0.771974	-0.426706	True
29	day_of_week_mon	-0.655685	-0.822846	-0.488523	True
30	day_of_week_thu	-0.638742	-0.803176	-0.474309	True
31	day_of_week_tue	-0.522563	-0.689353	-0.355772	True
32	day_of_week_wed	-0.513303	-0.679025	-0.347581	True
33	poutcome_failure	-0.677610	-0.948933	-0.406286	True

Summary for df_no_unknowns_resampled_feature_engineering (only statistically significant coefficients at 99%):

	Feature	Coefficient	Lower_CI_99	Upper_CI_99	Statistical_Significance_99
0	age	0.018981	0.013858	0.024105	True
1	duration	0.006293	0.006107	0.006480	True
2	pdays	-0.001663	-0.002096	-0.001230	True
3	cons.conf.idx	0.038864	0.007923	0.069806	True
4	nr.employed	-0.006393	-0.012641	-0.000145	True

Summary for df_no_unknowns_resampled_feature_engineering_no_duration (only statistically significant coefficients at 99%):

	Feature	Coefficient	Lower_CI_99	Upper_CI_99	Statistical_Significance_99
0	age	0.012920	0.008853	0.016988	True
1	campaign	-0.036754	-0.050174	-0.023334	True
2	pdays	-0.001697	-0.002120	-0.001274	True
3	cons.conf.idx	0.055402	0.029058	0.081746	True

Summary for df_no_unknowns_resampled_no_duration (only statistically significant coefficients at 99%):

	Feature	Coefficient	Lower_CI_99	Upper_CI_99	Statistical_Significance_99
0	age	0.046635	0.043257	0.050014	True
1	campaign	-0.255676	-0.275702	-0.235651	True
2	pdays	-0.001400	-0.001769	-0.001031	True
3	previous	-0.229708	-0.348844	-0.110572	True
4	cons.price.idx	0.271380	0.015964	0.526797	True
5	nr.employed	-0.004683	-0.007879	-0.001487	True
6	job_admin.	-0.252137	-0.353785	-0.150489	True
7	job_blue-collar	-0.249522	-0.383021	-0.116023	True
8	job_technician	-0.189143	-0.313191	-0.065095	True
9	marital_divorced	-0.145450	-0.275092	-0.015809	True
10	marital_married	-0.340208	-0.428064	-0.252352	True
11	marital_single	-0.221676	-0.314345	-0.129008	True
12	education_basic.9y	-0.189314	-0.323809	-0.054820	True
13	education_high.school	-0.286653	-0.394922	-0.178383	True
14	education_professional.course	-0.167414	-0.299737	-0.035090	True
15	education_university.degree	-0.235186	-0.330266	-0.140106	True
16	housing_no	-0.343836	-0.432672	-0.255000	True
17	housing_yes	-0.373139	-0.458384	-0.287893	True
18	loan_no	-0.174102	-0.271779	-0.076425	True
19	loan_yes	-0.190258	-0.319388	-0.061128	True
20	contact_telephone	-0.297716	-0.432897	-0.162535	True
21	month_may	-0.423994	-0.530205	-0.317783	True
22	day_of_week_fri	-0.231172	-0.344039	-0.118306	True
23	day_of_week_mon	-0.268705	-0.379878	-0.157531	True
24	day_of_week_thu	-0.224573	-0.332452	-0.116694	True
25	day_of_week_tue	-0.211121	-0.319740	-0.102502	True
26	day_of_week_wed	-0.192535	-0.300340	-0.084731	True
27	poutcome_failure	-0.216858	-0.401518	-0.032197	True

Summary for df_resampled (only statistically significant coefficients at 99%):

	Feature	Coefficient	Lower_CI_99	Upper_CI_99	Statistical_Significance_99
0	duration	0.006114	0.005935	0.006292	True
1	campaign	-0.579060	-0.609106	-0.549013	True
2	pdays	-0.000683	-0.001083	-0.000282	True
3	previous	-0.605868	-0.726436	-0.485301	True
4	emp.var.rate	-0.317158	-0.521315	-0.113000	True
5	cons.price.idx	0.464708	0.129434	0.799983	True
6	cons.conf.idx	0.154863	0.141701	0.168024	True
7	nr.employed	-0.006668	-0.010766	-0.002570	True
8	job_admin.	-0.461768	-0.595210	-0.328327	True
9	job_blue-collar	-0.698589	-0.862812	-0.534367	True
10	job_management	-0.216331	-0.413681	-0.018981	True
11	job_services	-0.319801	-0.512617	-0.126985	True
12	job_technician	-0.362994	-0.521756	-0.204232	True
13	marital_divorced	-0.309015	-0.477384	-0.140646	True
14	marital_married	-0.785581	-0.902292	-0.668871	True
15	marital_single	-0.375527	-0.502260	-0.248794	True
16	education_basic.4y	-0.272616	-0.463132	-0.082100	True
17	education_basic.9y	-0.476444	-0.651364	-0.301523	True
18	education_high.school	-0.631427	-0.775874	-0.486980	True
19	education_professional.course	-0.324938	-0.496905	-0.152972	True
20	education_university.degree	-0.362018	-0.488874	-0.235162	True
21	default_unknown	-0.517827	-0.717098	-0.318556	True
22	housing_no	-0.837712	-0.956064	-0.719361	True
23	housing_yes	-0.771201	-0.884094	-0.658308	True
24	loan_no	-0.354971	-0.483929	-0.226012	True
25	loan_yes	-0.437059	-0.608821	-0.265298	True
26	contact_telephone	-0.568498	-0.758189	-0.378807	True
27	month_may	-1.220764	-1.359075	-1.082453	True
28	month_nov	-0.203742	-0.389630	-0.017855	True
29	day_of_week_fri	-0.505946	-0.648394	-0.363497	True
30	day_of_week_mon	-0.563379	-0.702427	-0.424330	True
31	day_of_week_thu	-0.558241	-0.695080	-0.421402	True
32	day_of_week_tue	-0.472682	-0.612951	-0.332414	True
33	day_of_week_wed	-0.448493	-0.587040	-0.309947	True
34	poutcome_failure	-0.552897	-0.780649	-0.325145	True

Summary for df_resampled_feature_engineering (only statistically significant coefficients at 99%):

	Feature	Coefficient	Lower_CI_99	Upper_CI_99	Statistical_Significance_99
0	age	0.006331	0.002136	0.010526	True
1	duration	0.005885	0.005735	0.006035	True
2	campaign	-0.025030	-0.041426	-0.008635	True
3	pdays	-0.001532	-0.001905	-0.001159	True
4	cons.conf.idx	0.073199	0.046386	0.100013	True

Summary for df_resampled_feature_engineering_no_duration (only statistically significant coefficients at 99%):

	Feature	Coefficient	Lower_CI_99	Upper_CI_99	Statistical_Significance_99
0	age	-0.022821	-0.026291	-0.019351	True
1	campaign	-0.090854	-0.103774	-0.077935	True
2	pdays	-0.001665	-0.002057	-0.001272	True
3	cons.conf.idx	0.095519	0.070986	0.120052	True

Summary for df_resampled_no_duration (only statistically significant coefficients at 99%):

	Feature	Coefficient	Lower_CI_99	Upper_CI_99	Statistical_Significance_99
0	age	-0.005138	-0.008938	-0.001339	True
1	campaign	-0.834752	-0.867210	-0.802295	True
2	pdays	-0.002074	-0.002552	-0.001596	True
3	previous	-0.788825	-0.933630	-0.644019	True
4	cons.price.idx	0.710717	0.376512	1.044921	True
5	nr.employed	-0.010966	-0.015184	-0.006748	True
6	job_admin.	-0.900600	-1.029658	-0.771543	True
7	job_blue-collar	-1.026511	-1.172592	-0.880430	True
8	job_management	-0.360828	-0.540582	-0.181073	True
9	job_services	-0.502792	-0.674844	-0.330740	True
10	job_technician	-0.678416	-0.828718	-0.528115	True
11	marital_divorced	-0.518220	-0.673844	-0.362595	True
12	marital_married	-1.249270	-1.365674	-1.132865	True
13	marital_single	-0.778912	-0.905580	-0.652243	True
14	education_basic.4y	-0.455330	-0.628756	-0.281904	True
15	education_basic.6y	-0.285603	-0.489010	-0.082196	True
16	education_basic.9y	-0.724389	-0.880033	-0.568745	True
17	education_high.school	-0.984910	-1.119974	-0.849845	True
18	education_professional.course	-0.586373	-0.746156	-0.426589	True
19	education_university.degree	-0.852441	-0.977034	-0.727847	True
20	default_no	-0.362247	-0.505191	-0.219303	True
21	default_unknown	-0.861926	-1.030797	-0.693055	True
22	housing_no	-1.279704	-1.399230	-1.160178	True
23	housing_unknown	-0.135311	-0.267711	-0.002911	True
24	housing_yes	-1.368318	-1.484837	-1.251799	True
25	loan_no	-0.697764	-0.832153	-0.563374	True
26	loan_unknown	-0.135311	-0.267711	-0.002911	True
27	loan_yes	-0.700288	-0.866501	-0.534074	True
28	contact_cellular	-0.274394	-0.436167	-0.112621	True
29	contact_telephone	-1.141437	-1.321463	-0.961412	True
30	month_apr	-0.225956	-0.398576	-0.053337	True
31	month_aug	-0.350254	-0.529494	-0.171015	True
32	month_jun	-0.284416	-0.508193	-0.060640	True
33	month_may	-1.545404	-1.681846	-1.408962	True
34	month_nov	-0.392408	-0.568858	-0.215959	True
35	day_of_week_fri	-0.816923	-0.948786	-0.685059	True
36	day_of_week_mon	-0.988180	-1.119616	-0.856744	True
37	day_of_week_thu	-0.842070	-0.970046	-0.714093	True
38	day_of_week_tue	-0.815136	-0.945690	-0.684582	True
39	day_of_week_wed	-0.744801	-0.873776	-0.615826	True
40	poutcome_failure	-0.739909	-0.979757	-0.500062	True

Ap. Figure 16 (set) Coefficient Values of statistical significance, per tuned Logistic Regression model iteration.

Training Metrics Table

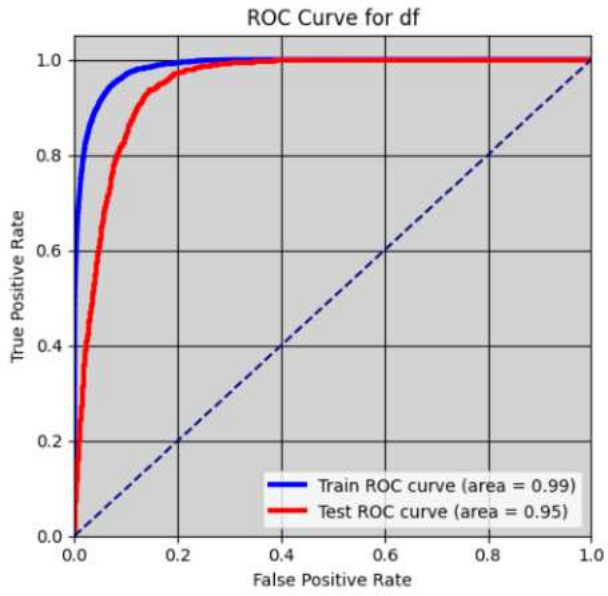
	DataFrame	Accuracy	AUC	Error Rate	Sensitivity	Specificity	Precision	Recall	Log Loss
0	df	0.960740	0.985492	0.039260	0.749798	0.987608	0.885152	0.749798	0.106919
1	df_no_duration	0.930376	0.906555	0.069624	0.438054	0.993085	0.889738	0.438054	0.205321
2	df_no_unknowns	0.960084	0.985914	0.039916	0.789793	0.985217	0.887455	0.789793	0.114864
3	df_no_unknowns_no_duration	0.925049	0.915282	0.074951	0.474322	0.991573	0.892557	0.474322	0.215204
4	df_resampled	0.976903	0.997938	0.023097	0.970059	0.983747	0.983521	0.970059	0.064314
5	df_resampled_no_duration	0.958374	0.987427	0.041626	0.927902	0.988845	0.988121	0.927902	0.122951
6	df_no_unknowns_resampled	0.978748	0.998126	0.021252	0.972338	0.985158	0.984967	0.972338	0.064788
7	df_no_unknowns_resampled_no_duration	0.956345	0.987661	0.043655	0.924858	0.987835	0.987019	0.924858	0.126482
8	df_feature_engineering	0.960466	0.985365	0.039534	0.757323	0.986342	0.875971	0.757323	0.107487
9	df_feature_engineering_no_duration	0.929678	0.903916	0.070322	0.431604	0.993120	0.888766	0.431604	0.208061
10	df_no_unknowns_feature_engineering	0.964350	0.989011	0.035650	0.809569	0.987195	0.903203	0.809569	0.107832
11	df_no_unknowns_feature_engineering_no_duration	0.926034	0.913044	0.073966	0.478150	0.992138	0.899760	0.478150	0.215672
12	df_resampled_feature_engineering	0.978802	0.998170	0.021198	0.972591	0.985013	0.984824	0.972591	0.062285
13	df_resampled_feature_engineering_no_duration	0.957724	0.987021	0.042276	0.926636	0.988811	0.988069	0.926636	0.124133
14	df_no_unknowns_resampled_feature_engineering	0.979640	0.998310	0.020360	0.973231	0.986050	0.985871	0.973231	0.063415
15	df_no_unknowns_resampled_feature_engineering_n...	0.957190	0.987441	0.042810	0.925046	0.989338	0.988607	0.925046	0.127023

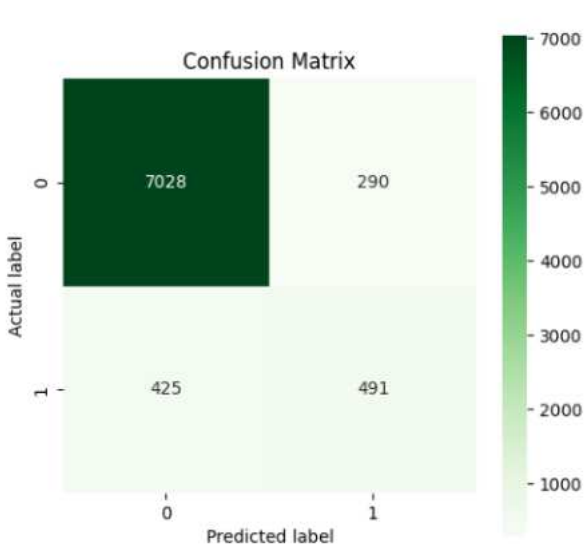
Testing Metrics Table

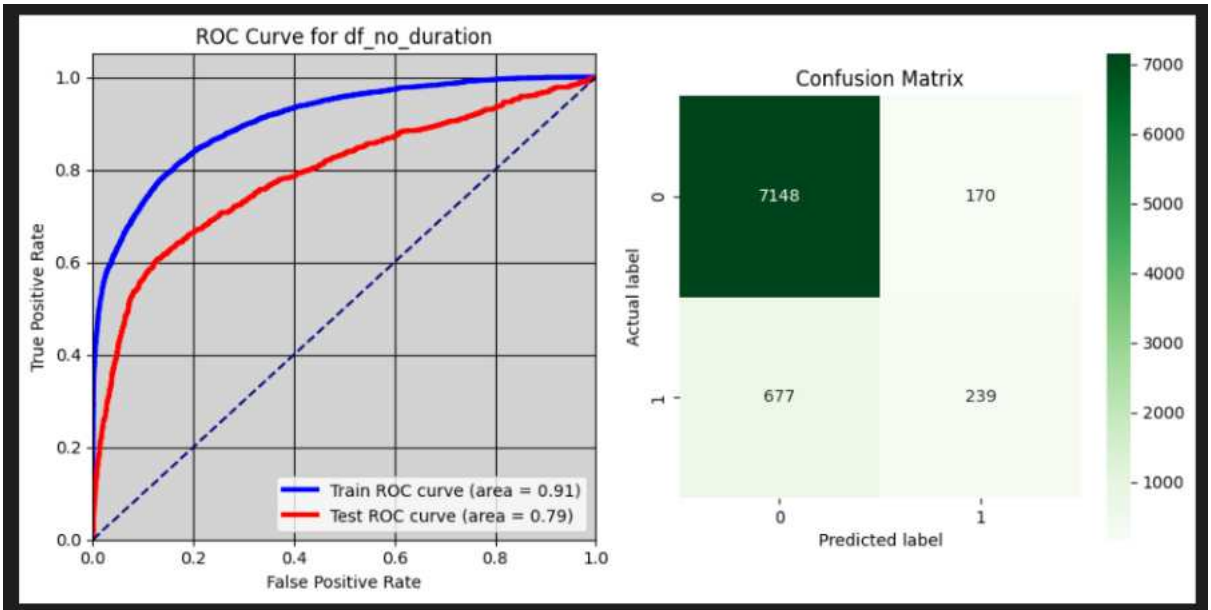
	DataFrame	Accuracy	AUC	Error Rate	Sensitivity	Specificity	Precision	Recall	Log Loss
0	df	0.913165	0.946568	0.086835	0.536026	0.960372	0.628681	0.536026	0.180789
1	df_no_duration	0.897134	0.786892	0.102866	0.260917	0.976770	0.584352	0.260917	0.283252
2	df_no_unknowns	0.911239	0.947684	0.088761	0.571429	0.956829	0.639752	0.571429	0.181721
3	df_no_unknowns_no_duration	0.890074	0.795706	0.109926	0.292649	0.970227	0.568733	0.292649	0.293021
4	df_resampled	0.951345	0.993171	0.048655	0.943616	0.959075	0.958438	0.943616	0.103820
5	df_resampled_no_duration	0.939848	0.971667	0.060152	0.907623	0.972078	0.970158	0.907623	0.165636
6	df_no_unknowns_resampled	0.945613	0.991774	0.054387	0.939121	0.952104	0.951456	0.939121	0.114053
7	df_no_unknowns_resampled_no_duration	0.932651	0.970588	0.067349	0.901541	0.963749	0.961330	0.901541	0.178803
8	df_feature_engineering	0.914987	0.947376	0.085013	0.548035	0.960918	0.637056	0.548035	0.177784
9	df_feature_engineering_no_duration	0.897012	0.787790	0.102988	0.268559	0.975676	0.580189	0.268559	0.281231
10	df_no_unknowns_feature_engineering	0.913372	0.948765	0.086628	0.572816	0.959062	0.652449	0.572816	0.179875
11	df_no_unknowns_feature_engineering_no_duration	0.891715	0.792406	0.108285	0.303745	0.970599	0.580902	0.303745	0.293783
12	df_resampled_feature_engineering	0.951345	0.993301	0.048655	0.943889	0.958801	0.958183	0.943889	0.102766
13	df_resampled_feature_engineering_no_duration	0.942722	0.972155	0.057278	0.910086	0.975363	0.973646	0.910086	0.164686
14	df_no_unknowns_resampled_feature_engineering	0.945238	0.991818	0.054762	0.939309	0.951165	0.950561	0.939309	0.114021
15	df_no_unknowns_resampled_feature_engineering_n...	0.933402	0.970909	0.066598	0.901541	0.965252	0.962874	0.901541	0.177423

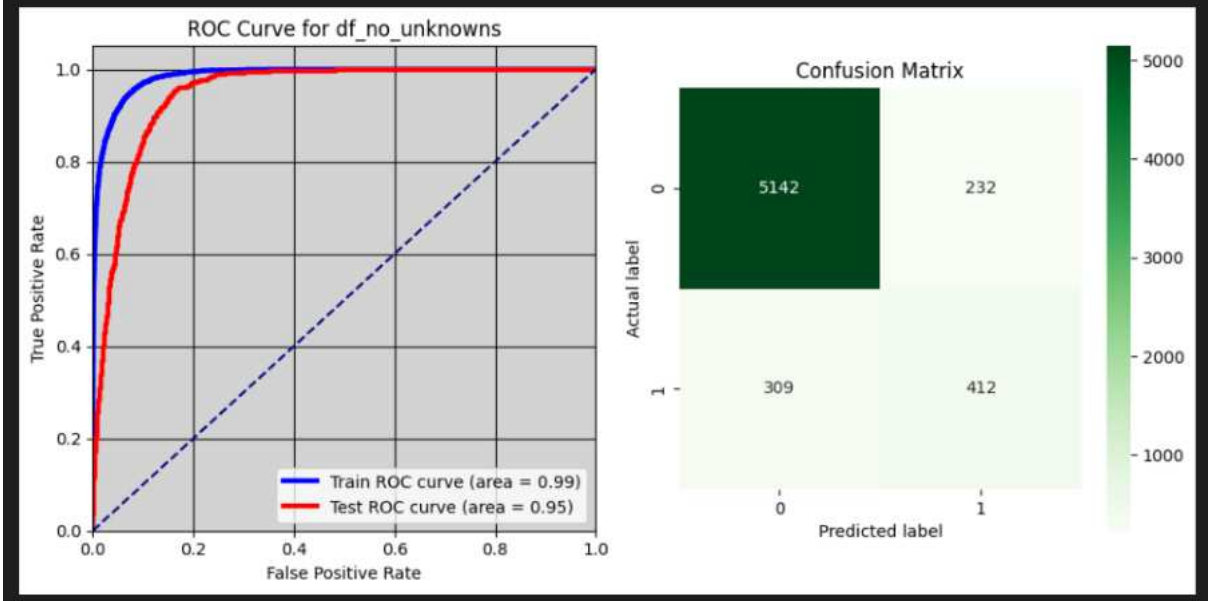
Ap. Figure 17 XGBoost Classification Performance metrics per data frame.

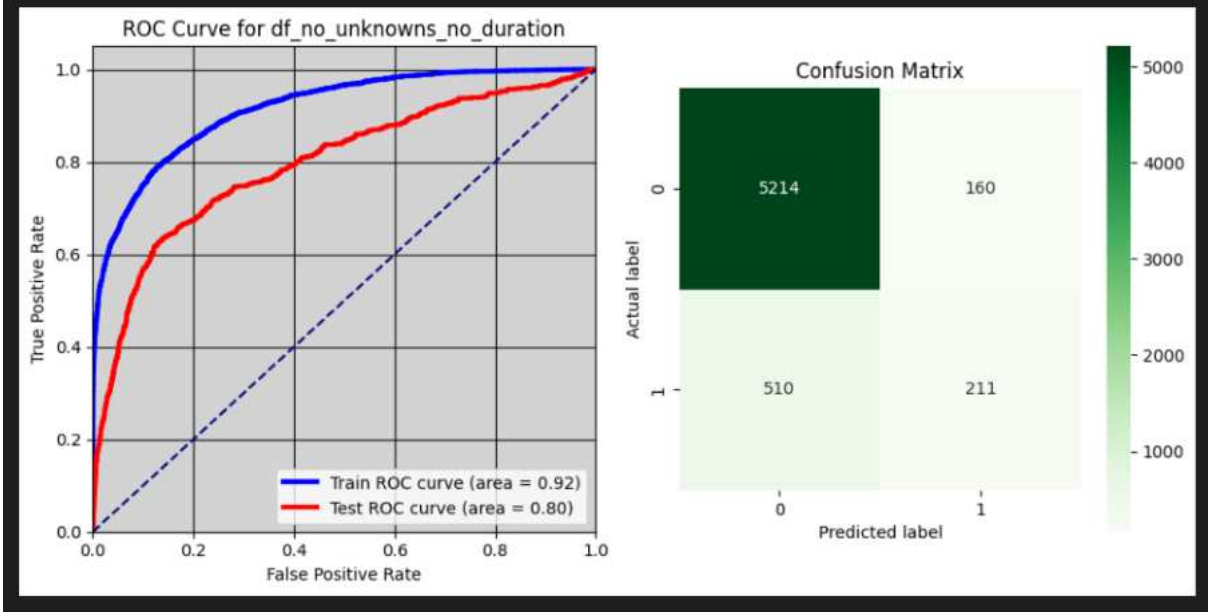
XGBoost Iterations performed per data frame.

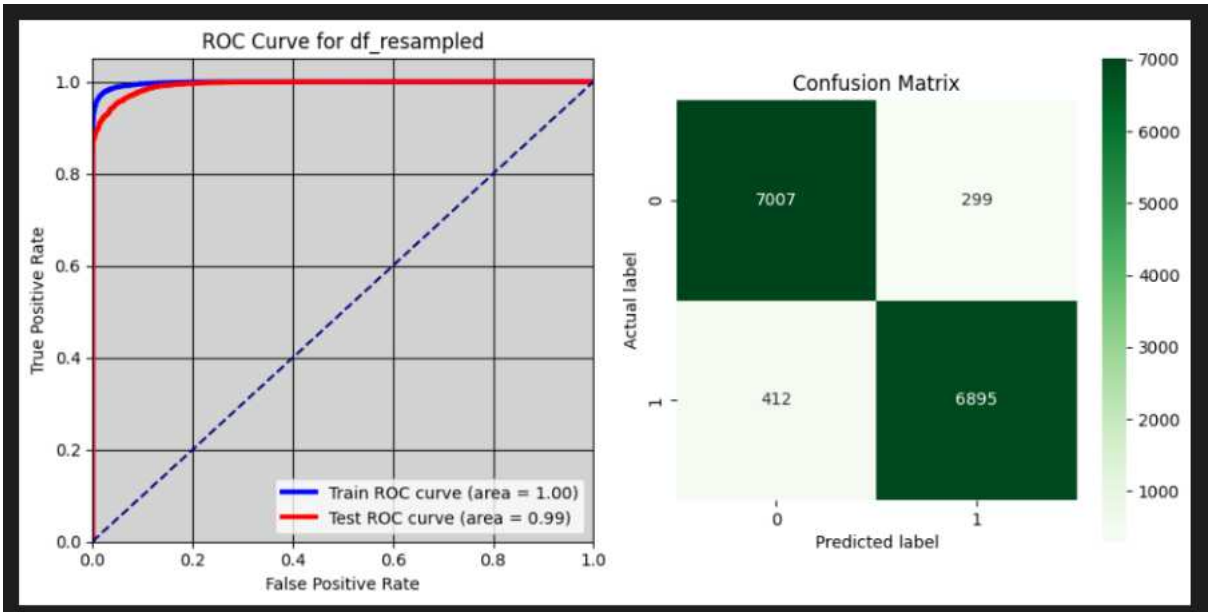


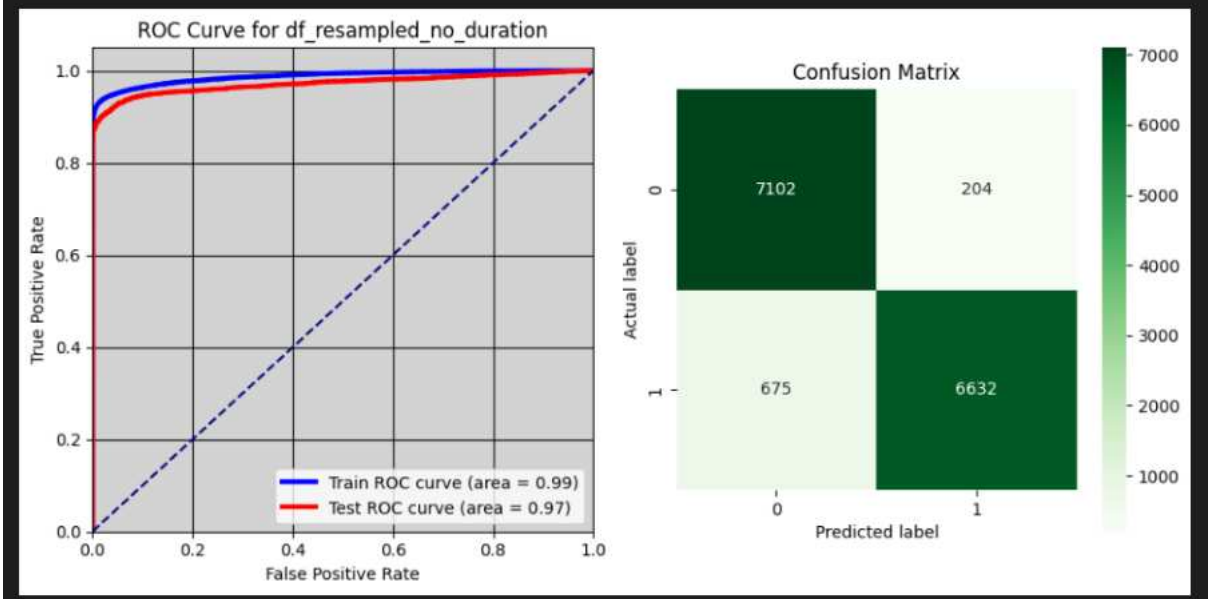


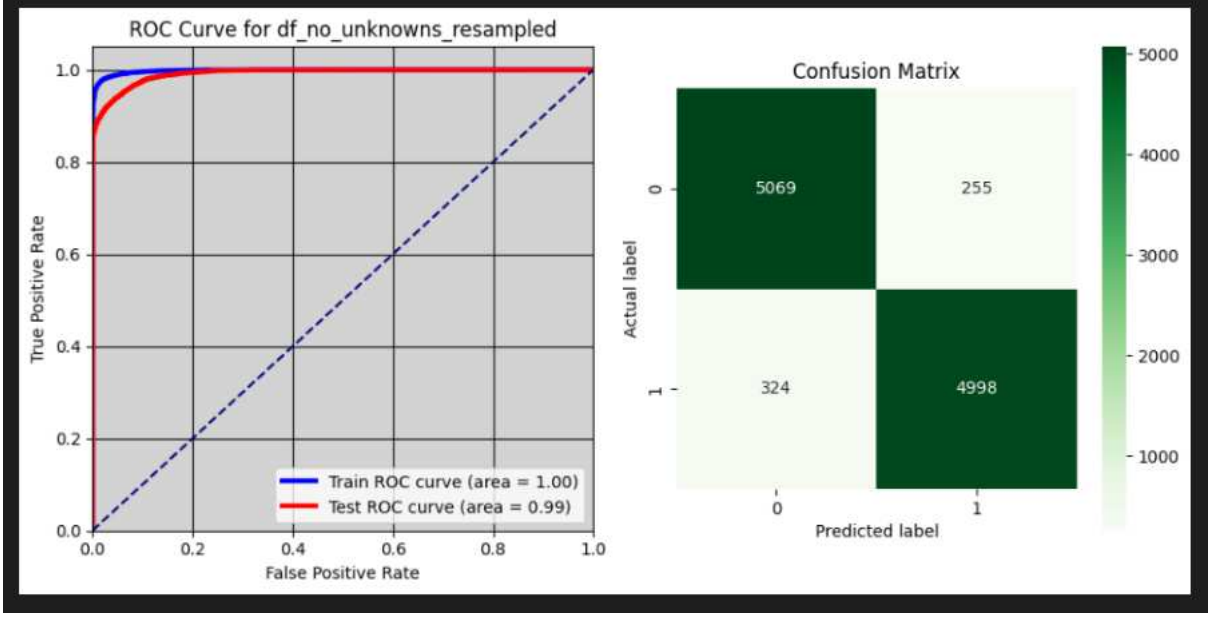

ROC Curve for df_no_duration
True Positive Rate
False Positive Rate
Train ROC curve (area = 0.91)
Test ROC curve (area = 0.79)
Confusion Matrix
Actual label
Predicted label
7148
170
677
239

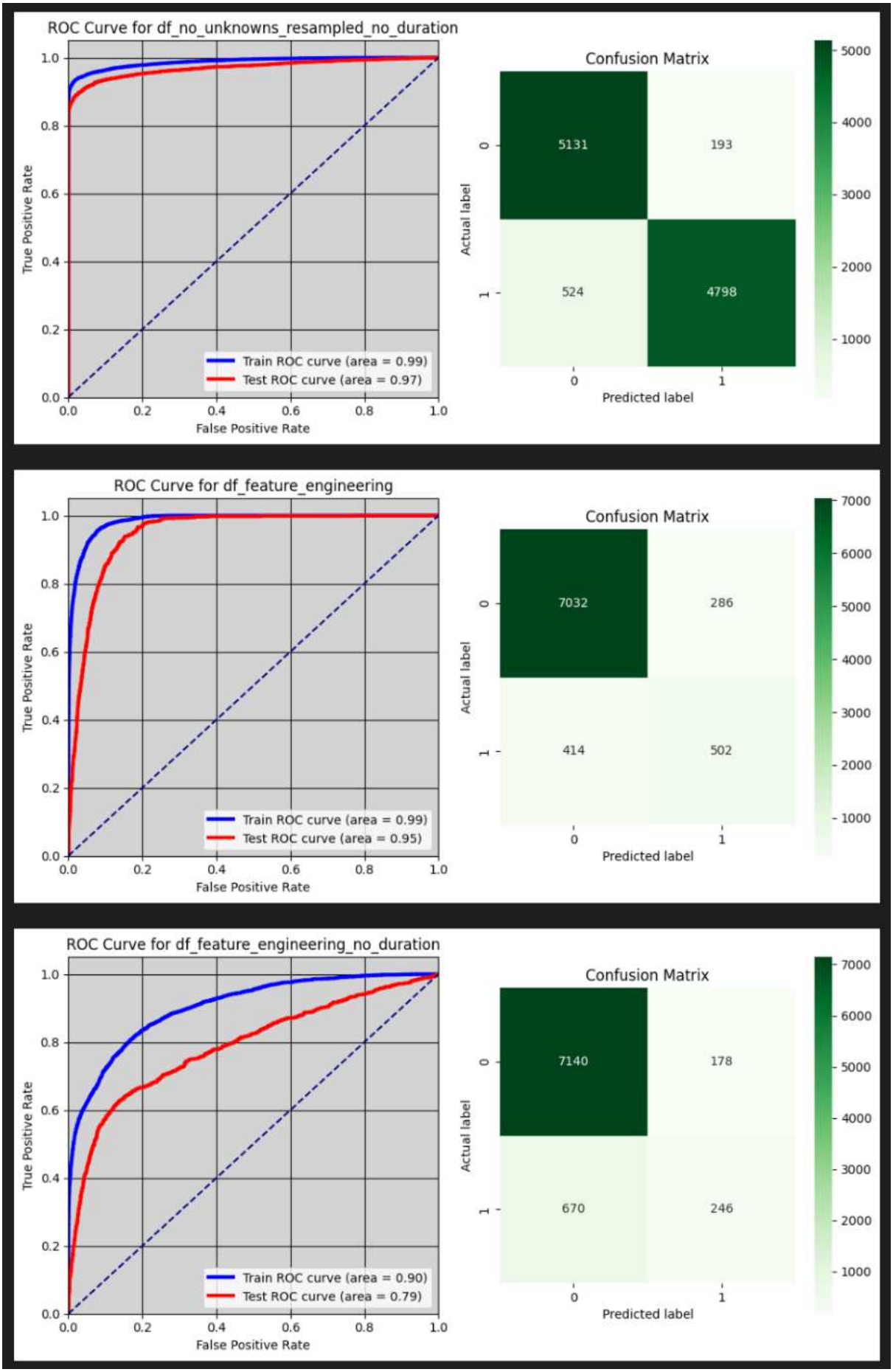

ROC Curve for df_no_unknowns
True Positive Rate
False Positive Rate
Train ROC curve (area = 0.99)
Test ROC curve (area = 0.95)
Confusion Matrix
Actual label
Predicted label
5142
232
309
412

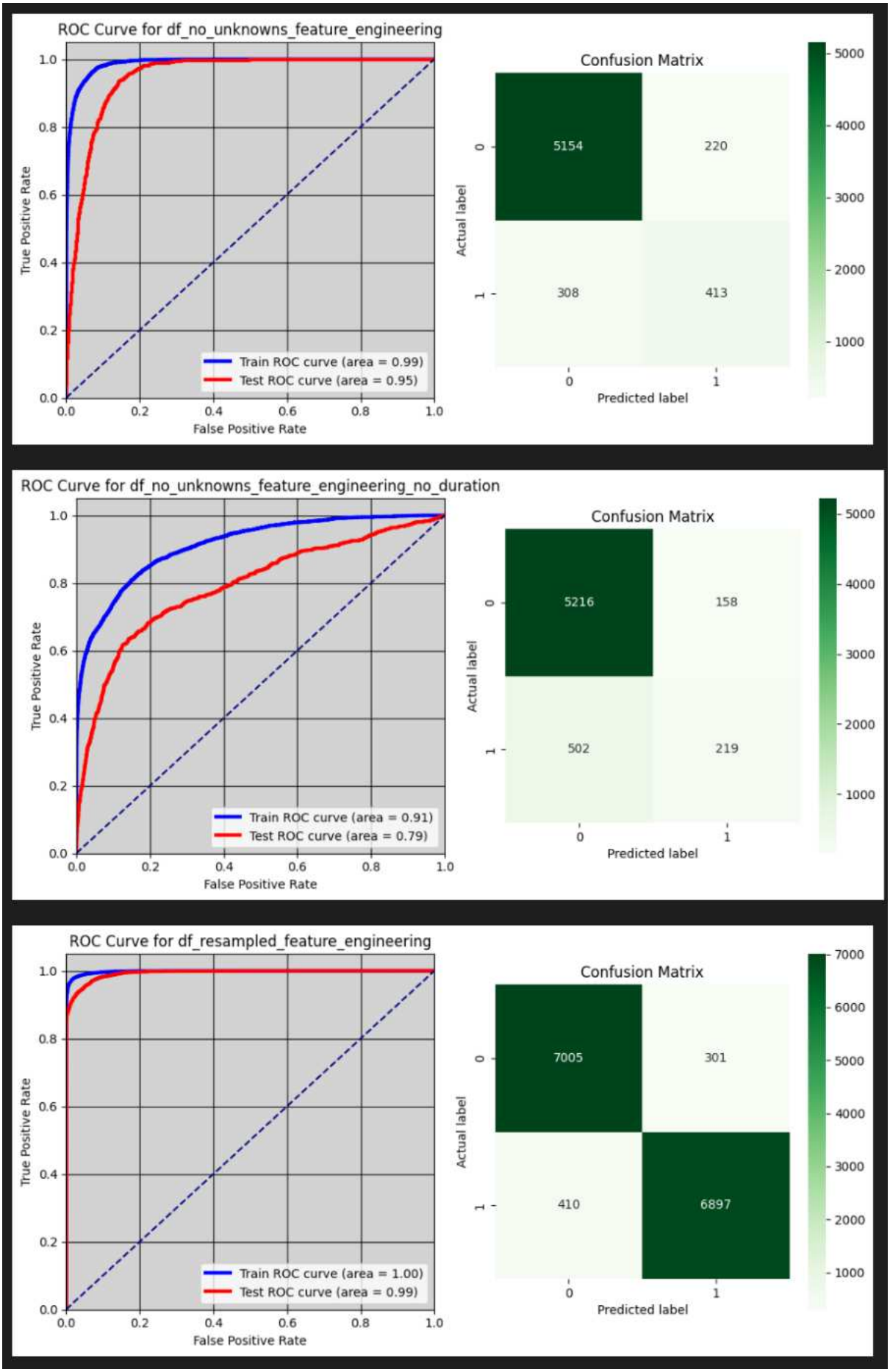

ROC Curve for df_no_unknowns_no_duration
True Positive Rate
False Positive Rate
Train ROC curve (area = 0.92)
Test ROC curve (area = 0.80)
Confusion Matrix
Actual label
Predicted label
5214
160
510
211

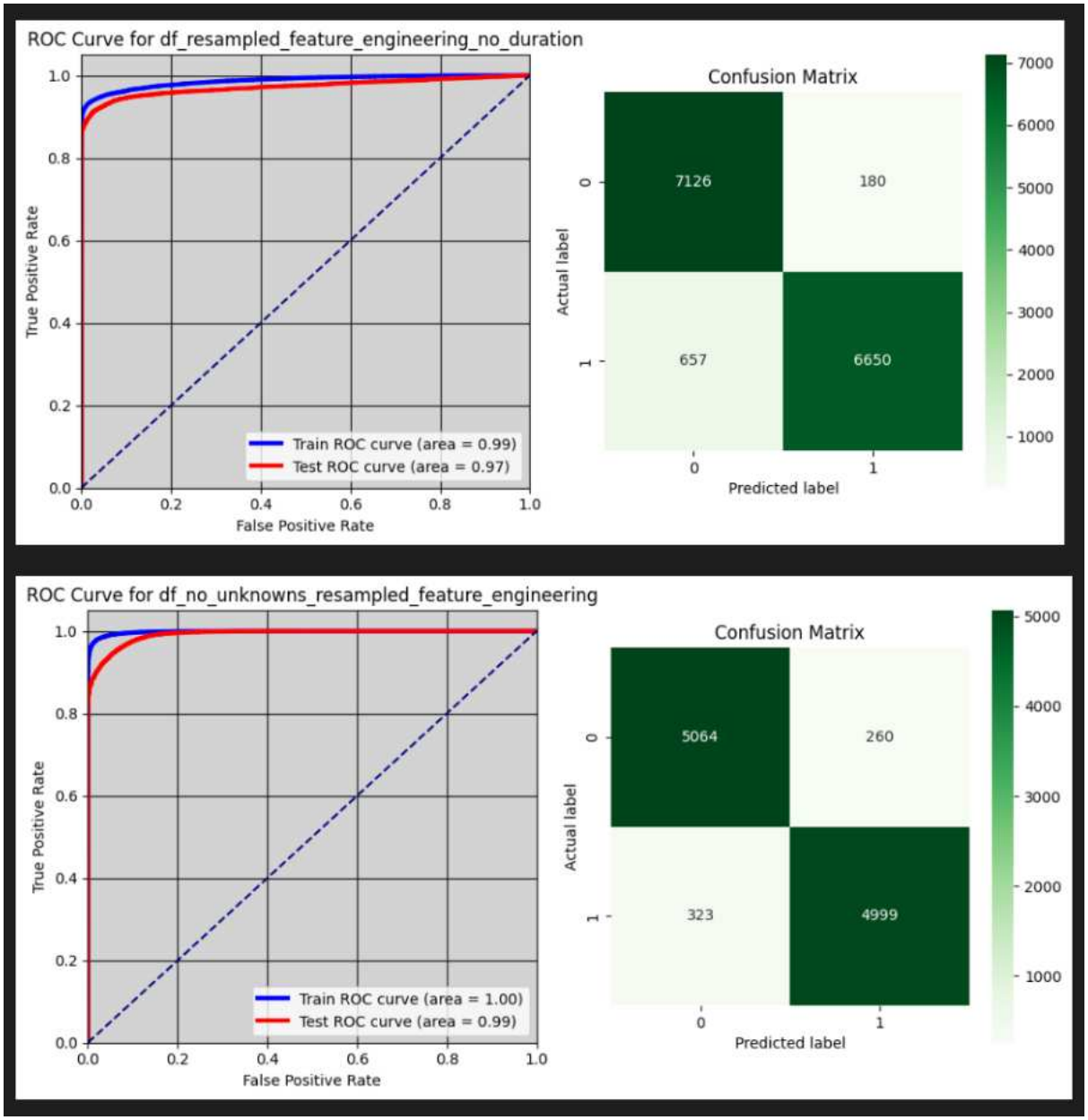

ROC Curve for df_resampled
True Positive Rate
False Positive Rate
Train ROC curve (area = 1.00)
Test ROC curve (area = 0.99)
Confusion Matrix
Actual label
Predicted label
7007
299
412
6895

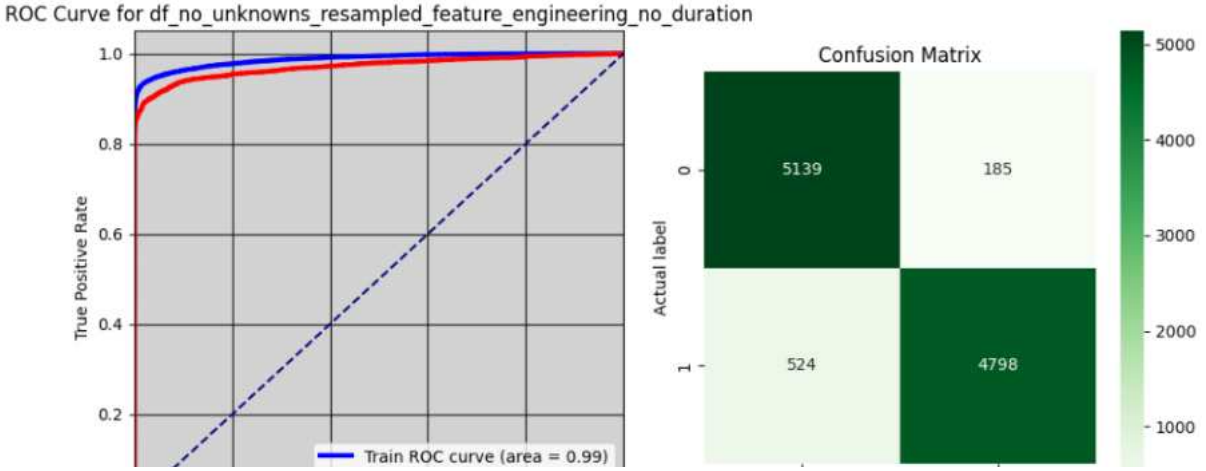

ROC Curve for df_resampled_no_duration
True Positive Rate
False Positive Rate
Train ROC curve (area = 0.99)
Test ROC curve (area = 0.97)
Confusion Matrix
Actual label
Predicted label
7102
204
675
6632


ROC Curve for df_no_unknowns_resampled
True Positive Rate
False Positive Rate
Train ROC curve (area = 1.00)
Test ROC curve (area = 0.99)
Confusion Matrix
Actual label
Predicted label
5069
255
324
4998


ROC Curve for df_no_unknowns_resampled_no_duration
True Positive Rate
False Positive Rate
Train ROC curve (area = 0.99)
Test ROC curve (area = 0.97)
Confusion Matrix
Actual label
Predicted label
5131
193
524
4798
ROC Curve for df_feature_engineering
Train ROC curve (area = 0.99)
Test ROC curve (area = 0.95)
Confusion Matrix
7032
286
414
502
ROC Curve for df_feature_engineering_no_duration
Train ROC curve (area = 0.90)
Test ROC curve (area = 0.79)
Confusion Matrix
7140
178
670
246


ROC Curve for df_no_unknowns_feature_engineering
True Positive Rate
False Positive Rate
Train ROC curve (area = 0.99)
Test ROC curve (area = 0.95)
Confusion Matrix
5154
220
308
413
Actual label
Predicted label
ROC Curve for df_no_unknowns_feature_engineering_no_duration
Train ROC curve (area = 0.91)
Test ROC curve (area = 0.79)
Confusion Matrix
5216
158
502
219
ROC Curve for df_resampled_feature_engineering
Train ROC curve (area = 1.00)
Test ROC curve (area = 0.99)
Confusion Matrix
7005
301
410
6897





Ap. Figure 17 (set) XGBoost Classification Model iterations performed per data frame, with corresponding Confusion matrix.

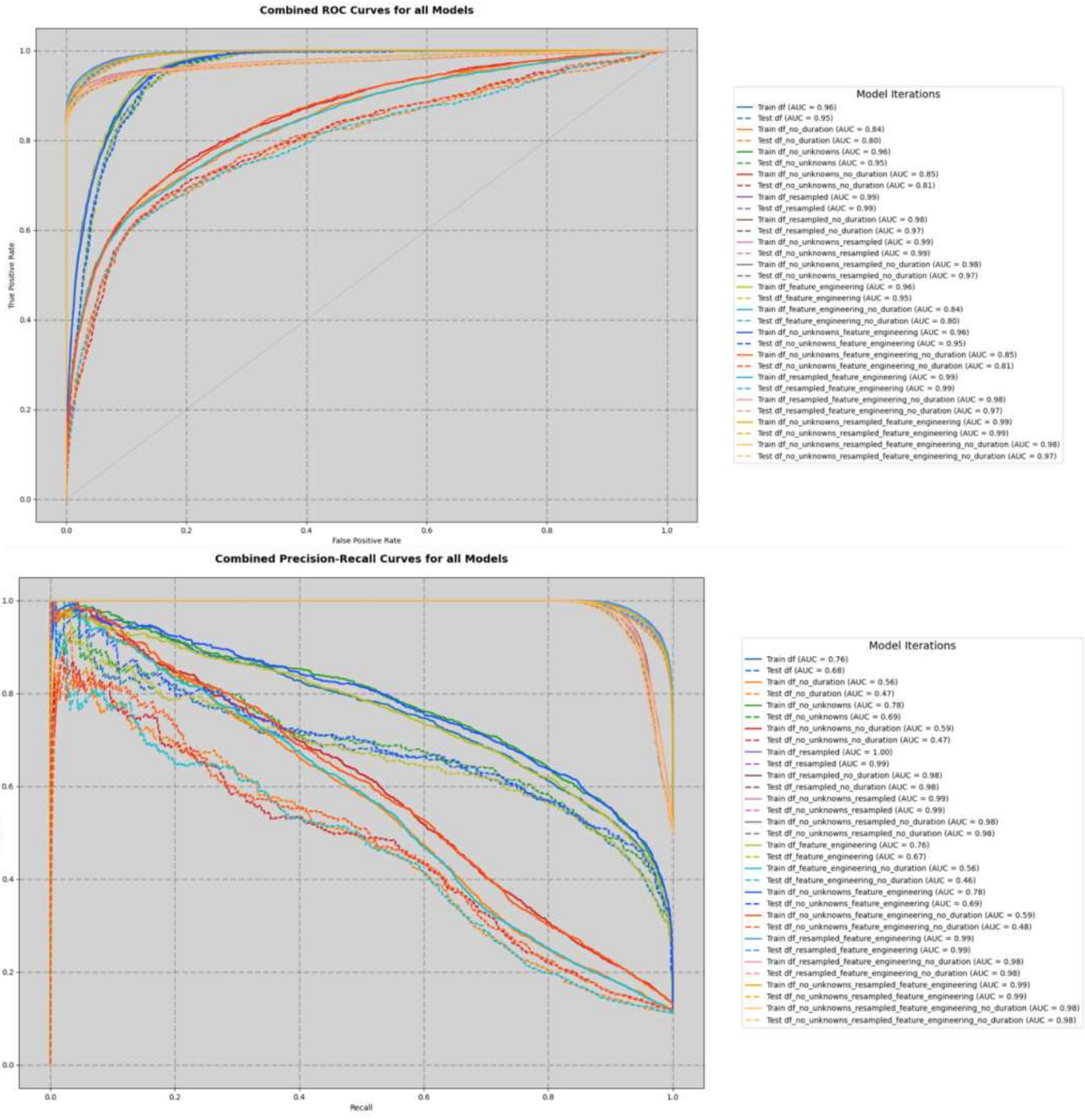
Training Metrics Table

	DataFrame	Accuracy	AUC	Error Rate	Sensitivity	Specificity	Precision	Recall	Log Loss
0	df	0.931833	0.962144	0.068167	0.609514	0.972889	0.741176	0.609514	0.152980
1	df_no_duration	0.909516	0.842180	0.090484	0.303144	0.986752	0.744554	0.303144	0.251580
2	df_no_unknowns	0.925418	0.960344	0.074582	0.631898	0.968740	0.748960	0.631898	0.167244
3	df_no_unknowns_no_duration	0.899943	0.854175	0.100057	0.334928	0.983334	0.747863	0.334928	0.267846
4	df_resampled	0.957878	0.994706	0.042122	0.953121	0.962635	0.962275	0.953121	0.096450
5	df_resampled_no_duration	0.945149	0.978155	0.054851	0.912127	0.978169	0.976625	0.912127	0.157974
6	df_no_unknowns_resampled	0.954279	0.993635	0.045721	0.950218	0.958339	0.958002	0.950218	0.105732
7	df_no_unknowns_resampled_no_duration	0.939296	0.976580	0.060704	0.903630	0.974966	0.973045	0.903630	0.167979
8	df_feature_engineering	0.930346	0.961931	0.069654	0.594464	0.973128	0.738071	0.594464	0.153539
9	df_feature_engineering_no_duration	0.910397	0.841120	0.089603	0.310132	0.986855	0.750325	0.310132	0.251643
10	df_no_unknowns_feature_engineering	0.924434	0.960393	0.075566	0.627751	0.968222	0.744608	0.627751	0.166944
11	df_no_unknowns_feature_engineering_no_duration	0.901132	0.853503	0.098868	0.345455	0.983146	0.751561	0.345455	0.268162
12	df_resampled_feature_engineering	0.957673	0.994530	0.042327	0.951204	0.964140	0.963669	0.951204	0.098211
13	df_resampled_feature_engineering_no_duration	0.945457	0.977937	0.054543	0.912367	0.978546	0.977025	0.912367	0.159674
14	df_no_unknowns_resampled_feature_engineering	0.954537	0.993506	0.045463	0.950077	0.958997	0.958631	0.950077	0.107274
15	df_no_unknowns_resampled_feature_engineering_n...	0.940095	0.975806	0.059905	0.902127	0.978066	0.976266	0.902127	0.170665

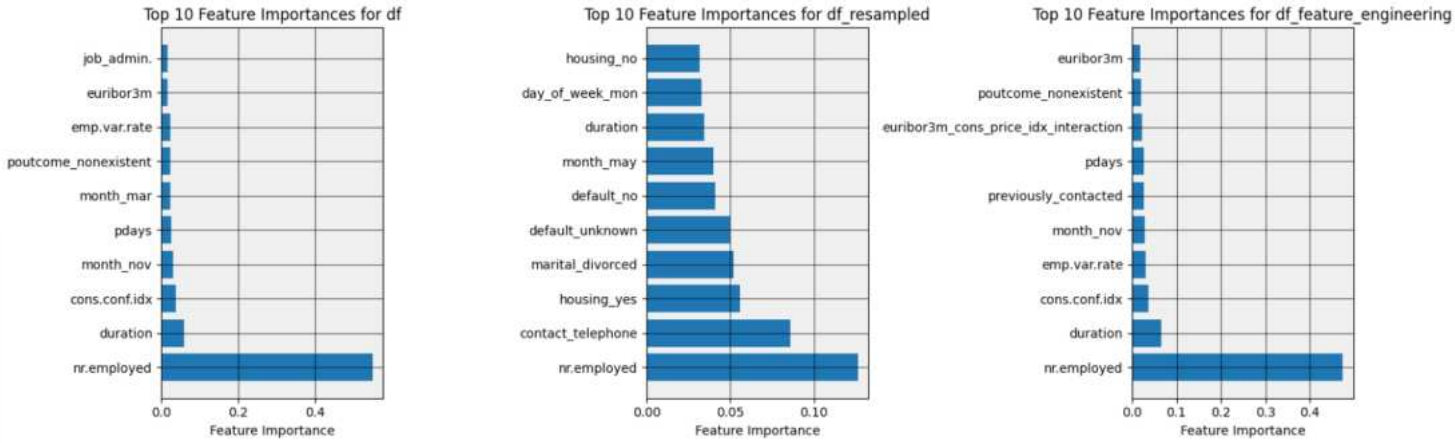
Testing Metrics Table

	DataFrame	Accuracy	AUC	Error Rate	Sensitivity	Specificity	Precision	Recall	Log Loss
0	df	0.919480	0.950672	0.080520	0.555677	0.965018	0.665359	0.555677	0.169603
1	df_no_duration	0.903328	0.798267	0.096672	0.255459	0.984422	0.672414	0.255459	0.270559
2	df_no_unknowns	0.917637	0.951113	0.082363	0.571429	0.964086	0.680992	0.571429	0.174702
3	df_no_unknowns_no_duration	0.894176	0.805450	0.105824	0.278779	0.976740	0.616564	0.278779	0.281164
4	df_resampled	0.949429	0.992798	0.050571	0.944437	0.954421	0.953967	0.944437	0.108997
5	df_resampled_no_duration	0.938958	0.972905	0.061042	0.906528	0.971393	0.969413	0.906528	0.170104
6	df_no_unknowns_resampled	0.947304	0.991968	0.052696	0.944194	0.950413	0.950085	0.944194	0.115783
7	df_no_unknowns_resampled_no_duration	0.934999	0.971900	0.065001	0.901165	0.968820	0.966546	0.901165	0.178069
8	df_feature_engineering	0.917537	0.949971	0.082463	0.550218	0.963515	0.653696	0.550218	0.170488
9	df_feature_engineering_no_duration	0.901506	0.796990	0.098494	0.254367	0.982509	0.645429	0.254367	0.271449
10	df_no_unknowns_feature_engineering	0.915176	0.951335	0.084824	0.568655	0.961667	0.665584	0.568655	0.174141
11	df_no_unknowns_feature_engineering_no_duration	0.893355	0.806471	0.106645	0.271845	0.976740	0.610592	0.271845	0.280022
12	df_resampled_feature_engineering	0.950934	0.992884	0.049066	0.943753	0.958117	0.957512	0.943753	0.108659
13	df_resampled_feature_engineering_no_duration	0.939780	0.972841	0.060220	0.905159	0.974405	0.972504	0.905159	0.170969
14	df_no_unknowns_resampled_feature_engineering	0.946647	0.991797	0.053353	0.942315	0.950977	0.950531	0.942315	0.117535
15	df_no_unknowns_resampled_feature_engineering_n...	0.933778	0.971021	0.066222	0.897031	0.970511	0.968161	0.897031	0.182378

Ap. Figure 18 Hyperparameter-tuned XGBoost Classification Performance metrics per data frame.



Ap. Figure 19 (set) Training and Testing Tuned XGBoost Classification Model iterations,
ROC curves and Precision-Recall curves – (params:: n_estimators: 64, min_child_weight: 2, max_deapth: 4, gamma: 0.2, alpha 0.9)



Ap. Figure 20 (set) XGBoost Feature importance(s) per key data frame.

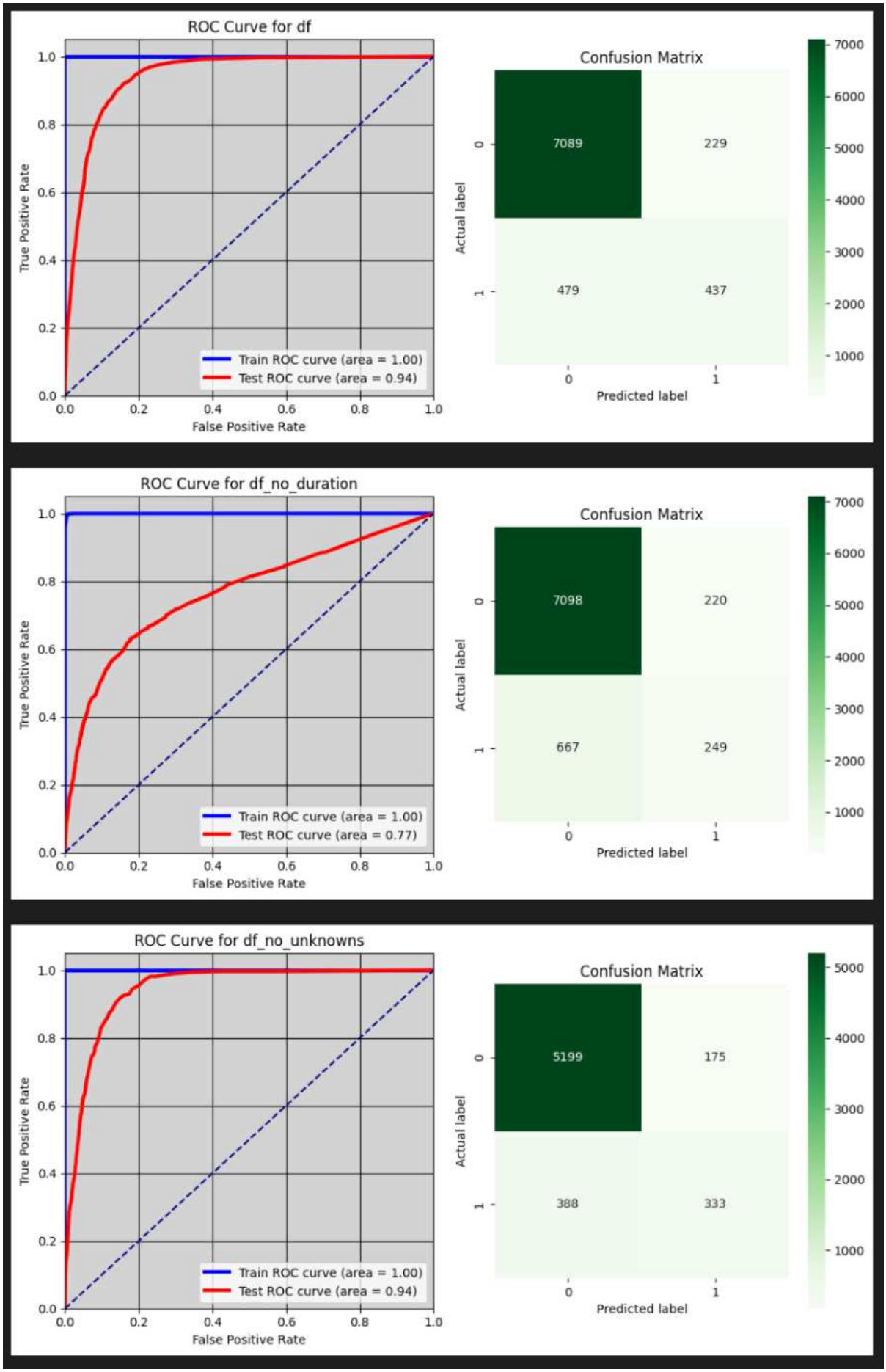
	DataFrame	Accuracy	AUC	Error Rate	Sensitivity	Specificity	Precision	Recall	Log Loss
0	df	0.999970	1.000000	0.000030	0.999731	1.000000	1.000000	0.999731	0.048035
1	df_no_duration	0.995020	0.999837	0.004980	0.963182	0.999076	0.992523	0.963182	0.067621
2	df_no_unknowns	0.999959	1.000000	0.000041	0.999681	1.000000	1.000000	0.999681	0.054184
3	df_no_unknowns_no_duration	0.993928	0.999793	0.006072	0.961404	0.998729	0.991121	0.961404	0.075244
4	df_resampled	1.000000	1.000000	0.000000	1.000000	1.000000	1.000000	1.000000	0.031402
5	df_resampled_no_duration	0.996989	0.999970	0.003011	0.995757	0.998221	0.998216	0.995757	0.044756
6	df_no_unknowns_resampled	1.000000	1.000000	0.000000	1.000000	1.000000	1.000000	1.000000	0.034932
7	df_no_unknowns_resampled_no_duration	0.996924	0.999966	0.003076	0.995257	0.998591	0.998586	0.995257	0.048778
8	df_feature_engineering	0.999970	1.000000	0.000030	0.999731	1.000000	1.000000	0.999731	0.048639
9	df_feature_engineering_no_duration	0.995051	0.999851	0.004949	0.962107	0.999247	0.993892	0.962107	0.067713
10	df_no_unknowns_feature_engineering	1.000000	1.000000	0.000000	1.000000	1.000000	1.000000	1.000000	0.054622
11	df_no_unknowns_feature_engineering_no_duration	0.994134	0.999793	0.005866	0.962998	0.998729	0.991136	0.962998	0.075459
12	df_resampled_feature_engineering	1.000000	1.000000	0.000000	1.000000	1.000000	1.000000	1.000000	0.032036
13	df_resampled_feature_engineering_no_duration	0.996920	0.999971	0.003080	0.995415	0.998426	0.998421	0.995415	0.044856
14	df_no_unknowns_resampled_feature_engineering	1.000000	1.000000	0.000000	1.000000	1.000000	1.000000	1.000000	0.035243
15	df_no_unknowns_resampled_feature_engineering_n...	0.996900	0.999962	0.003100	0.995351	0.998450	0.998445	0.995351	0.049285

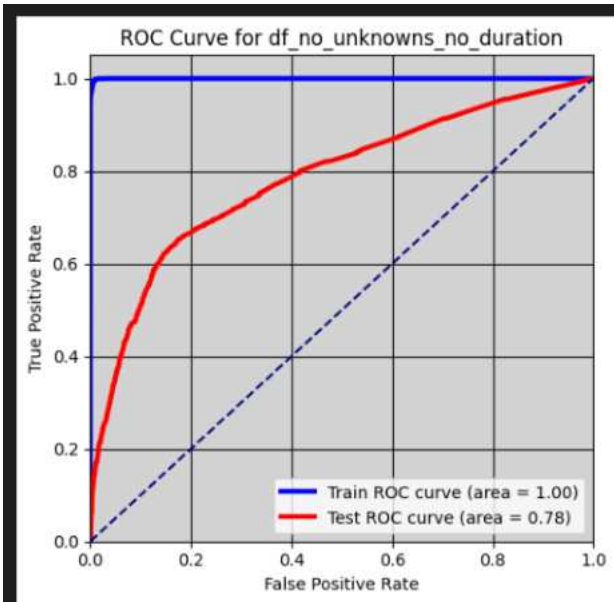
Testing Metrics Table

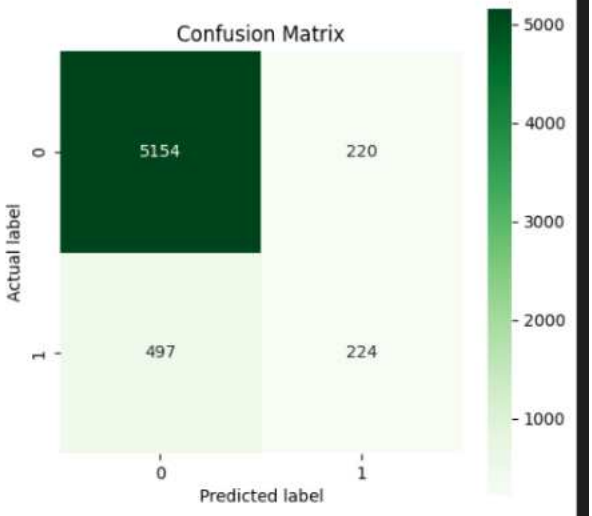
	DataFrame	Accuracy	AUC	Error Rate	Sensitivity	Specificity	Precision	Recall	Log Loss
0	df	0.912558	0.938911	0.087442	0.465066	0.968571	0.649390	0.465066	0.208332
1	df_no_duration	0.891912	0.765718	0.108088	0.280568	0.968434	0.526639	0.280568	0.496763
2	df_no_unknowns	0.909598	0.943183	0.090402	0.475728	0.967808	0.664729	0.475728	0.190572
3	df_no_unknowns_no_duration	0.884495	0.780182	0.115505	0.310680	0.961481	0.519722	0.310680	0.499240
4	df_resampled	0.953808	0.993043	0.046192	0.945395	0.962223	0.961581	0.945395	0.115834
5	df_resampled_no_duration	0.937111	0.972167	0.062889	0.917066	0.957158	0.955375	0.917066	0.256460
6	df_no_unknowns_resampled	0.949183	0.992074	0.050817	0.941188	0.957175	0.956464	0.941188	0.120672
7	df_no_unknowns_resampled_no_duration	0.931148	0.969021	0.068852	0.912251	0.950038	0.948057	0.912251	0.273588
8	df_feature_engineering	0.911708	0.938435	0.088292	0.453057	0.969117	0.647426	0.453057	0.202675
9	df_feature_engineering_no_duration	0.891790	0.768720	0.108210	0.271834	0.969391	0.526427	0.271834	0.522412
10	df_no_unknowns_feature_engineering	0.909598	0.942944	0.090402	0.466019	0.969111	0.669323	0.466019	0.191717
11	df_no_unknowns_feature_engineering_no_duration	0.884495	0.785302	0.115505	0.325936	0.959434	0.518764	0.325936	0.487620
12	df_resampled_feature_engineering	0.952508	0.993257	0.047492	0.942658	0.962360	0.961608	0.942658	0.114311
13	df_resampled_feature_engineering_no_duration	0.937864	0.973697	0.062136	0.920350	0.955379	0.953765	0.920350	0.224489
14	df_no_unknowns_resampled_feature_engineering	0.948713	0.992253	0.051287	0.941000	0.956424	0.955725	0.941000	0.120819
15	df_no_unknowns_resampled_feature_engineering_n...	0.929551	0.970083	0.070449	0.912251	0.946844	0.944920	0.912251	0.271763

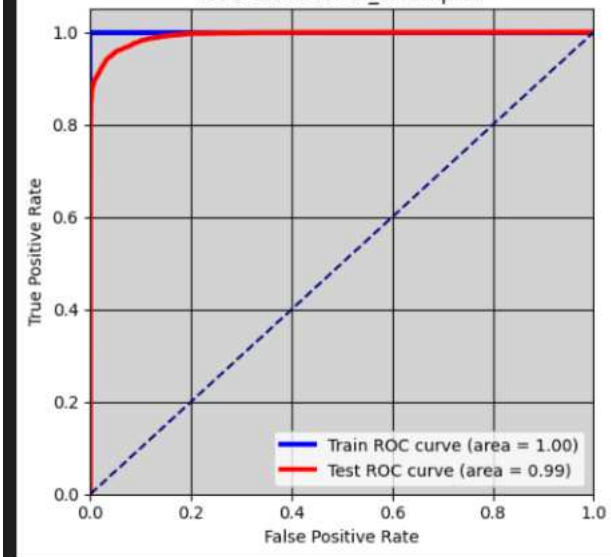
Ap. Figure 20 Random Forest Classification Performance metrics per data frame.

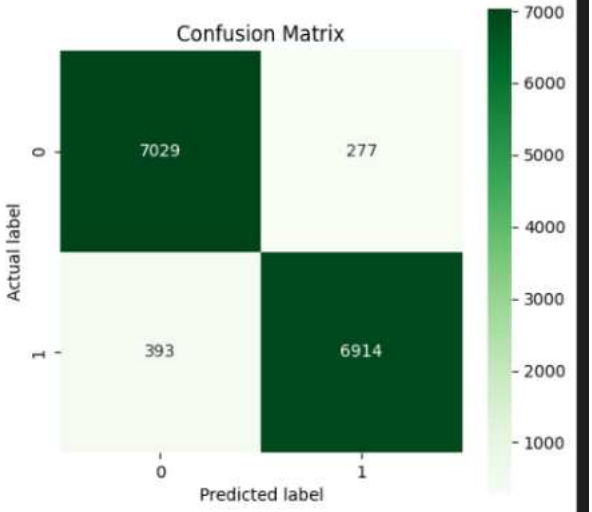
Random Forest Iterations performed per data frame.

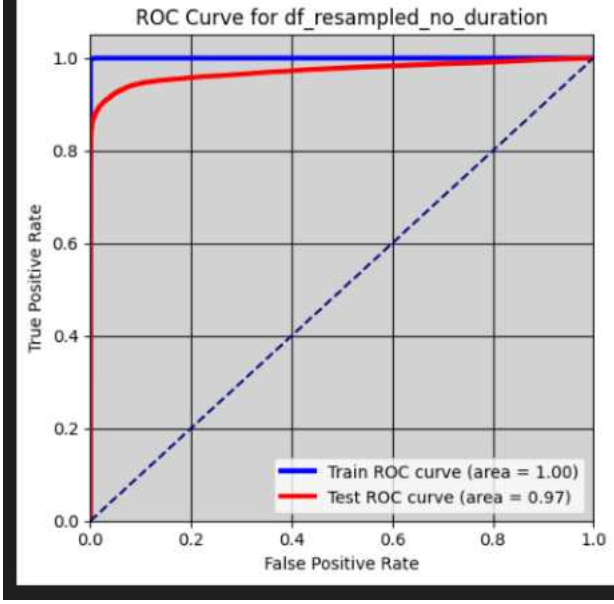


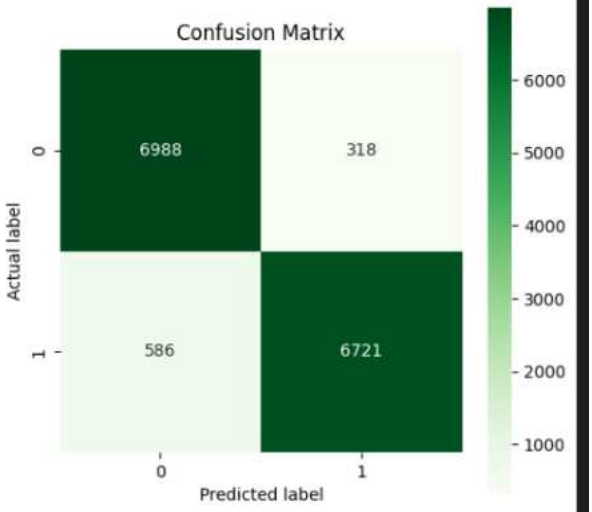

ROC Curve for df_no_unknowns_no_duration
True Positive Rate
False Positive Rate
Train ROC curve (area = 1.00)
Test ROC curve (area = 0.78)

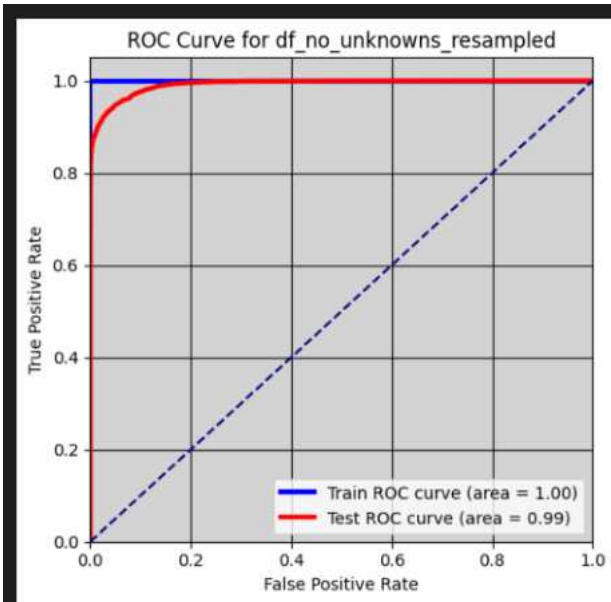

Confusion Matrix
5154
220
497
224
Actual label
Predicted label

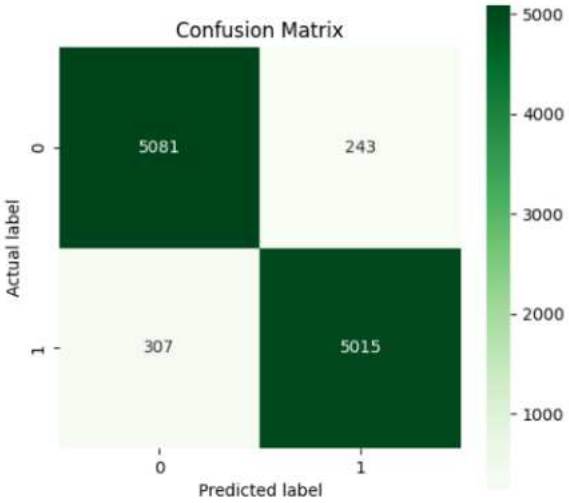

ROC Curve for df_resampled
True Positive Rate
False Positive Rate
Train ROC curve (area = 1.00)
Test ROC curve (area = 0.99)

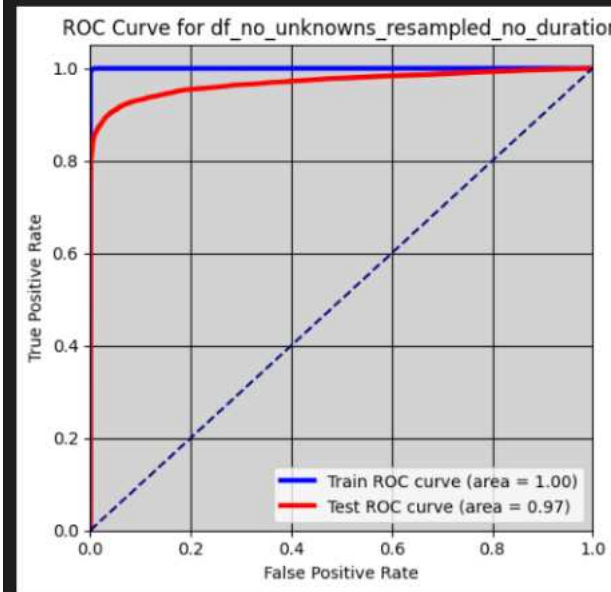

Confusion Matrix
7029
277
393
6914
Actual label
Predicted label

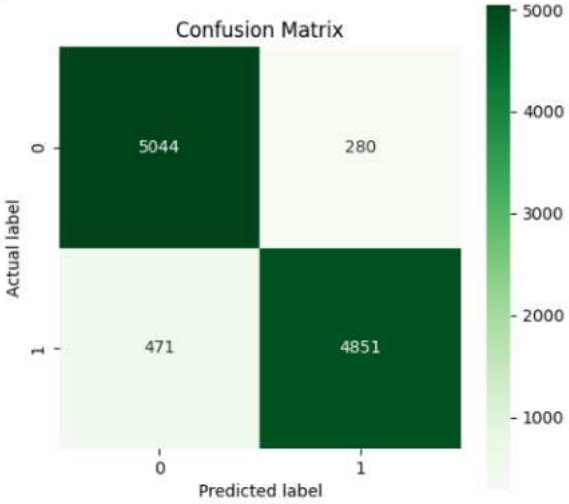

ROC Curve for df_resampled_no_duration
True Positive Rate
False Positive Rate
Train ROC curve (area = 1.00)
Test ROC curve (area = 0.97)

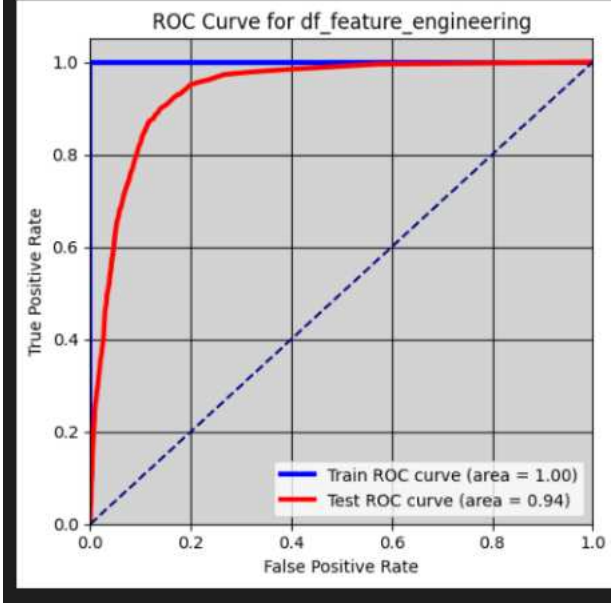

Confusion Matrix
6988
318
586
6721
Actual label
Predicted label

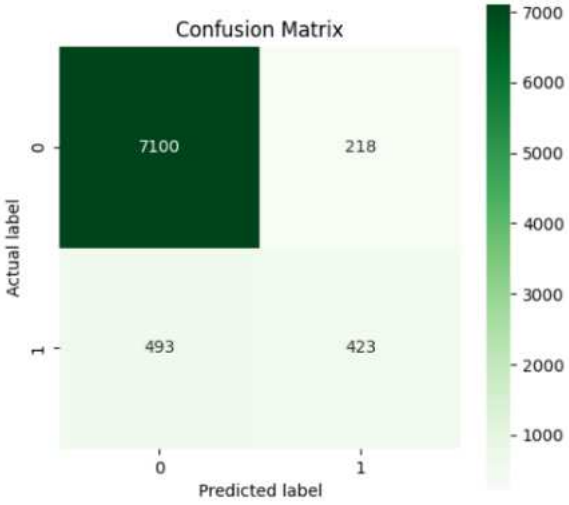

ROC Curve for df_no_unknowns_resampled
True Positive Rate
False Positive Rate
Train ROC curve (area = 1.00)
Test ROC curve (area = 0.99)

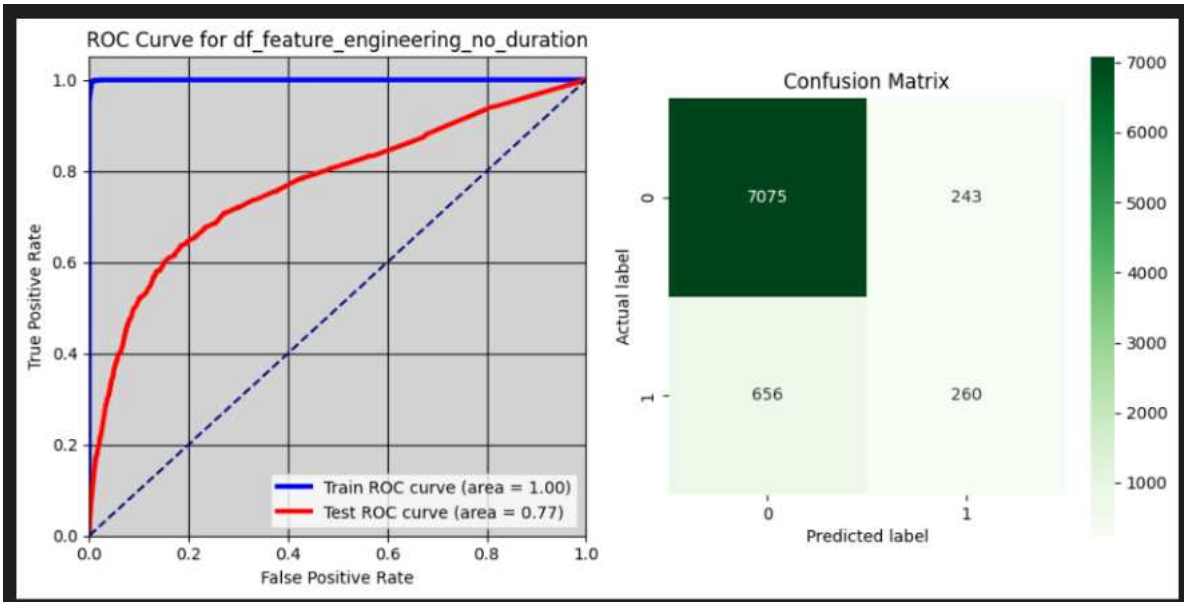

Confusion Matrix
5081
243
307
5015
Actual label
Predicted label

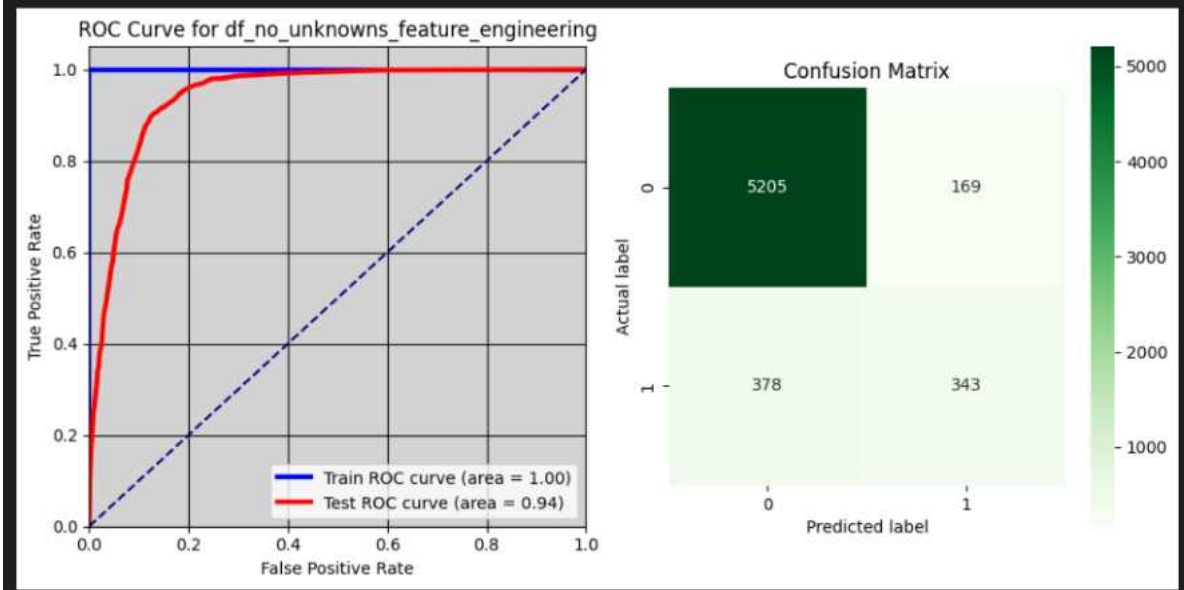

ROC Curve for df_no_unknowns_resampled_no_duration
True Positive Rate
False Positive Rate
Train ROC curve (area = 1.00)
Test ROC curve (area = 0.97)

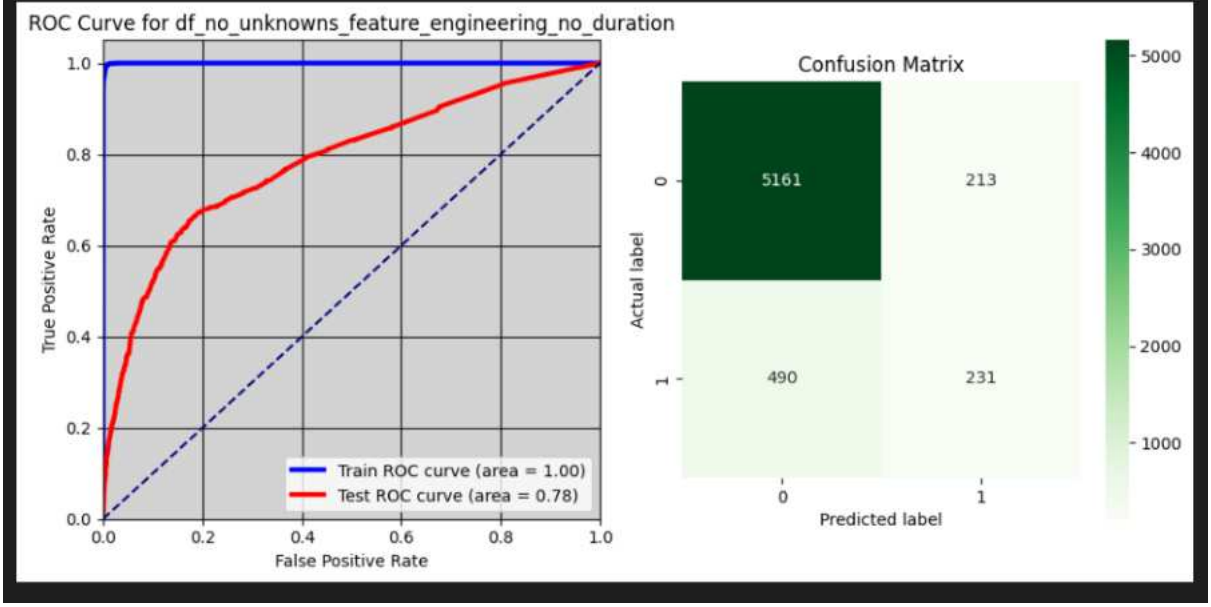

Confusion Matrix
5044
280
471
4851
Actual label
Predicted label

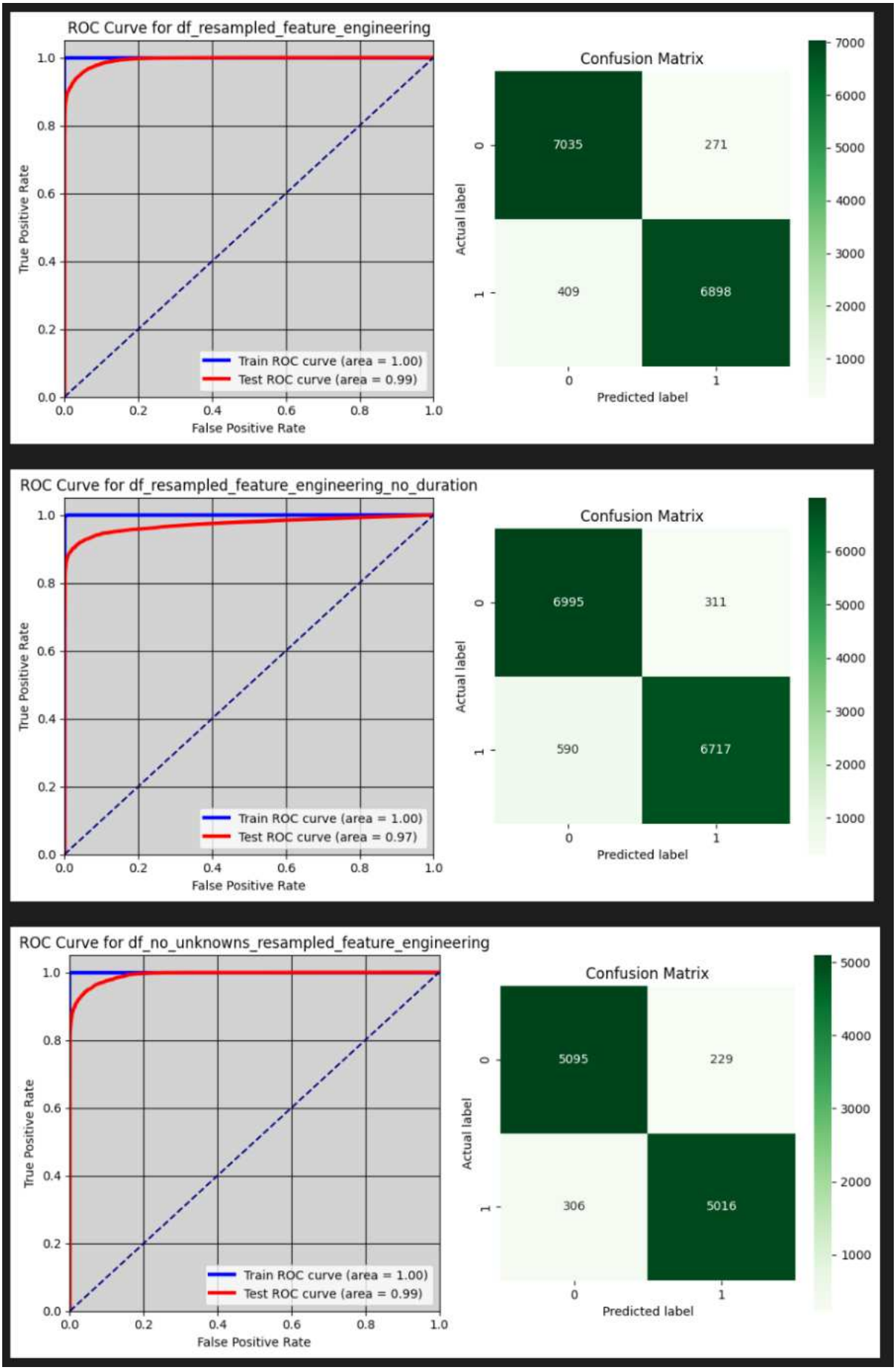

ROC Curve for df_feature_engineering
True Positive Rate
False Positive Rate
Train ROC curve (area = 1.00)
Test ROC curve (area = 0.94)

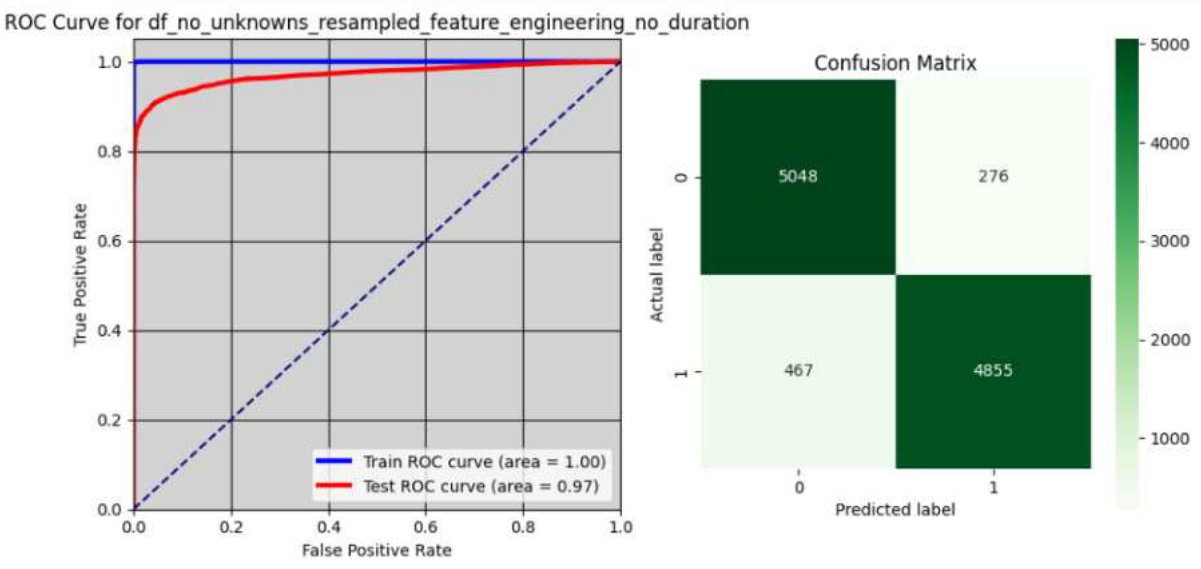

Confusion Matrix
7100
218
493
423
Actual label
Predicted label


ROC Curve for df_feature_engineering_no_duration
True Positive Rate
False Positive Rate
Train ROC curve (area = 1.00)
Test ROC curve (area = 0.77)
Confusion Matrix
Actual label
Predicted label
7075
243
656
260

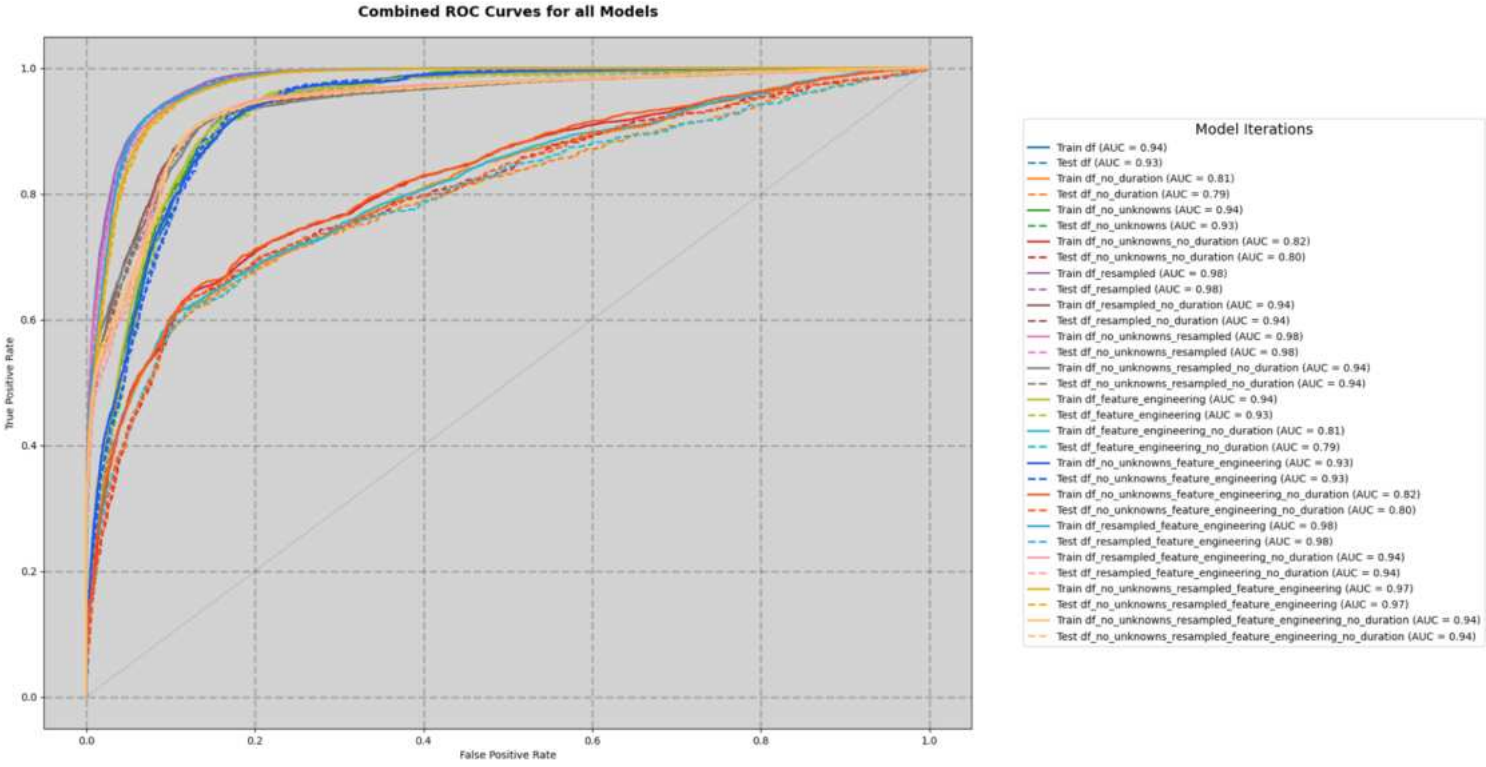

ROC Curve for df_no_unknowns_feature_engineering
True Positive Rate
False Positive Rate
Train ROC curve (area = 1.00)
Test ROC curve (area = 0.94)
Confusion Matrix
Actual label
Predicted label
5205
169
378
343

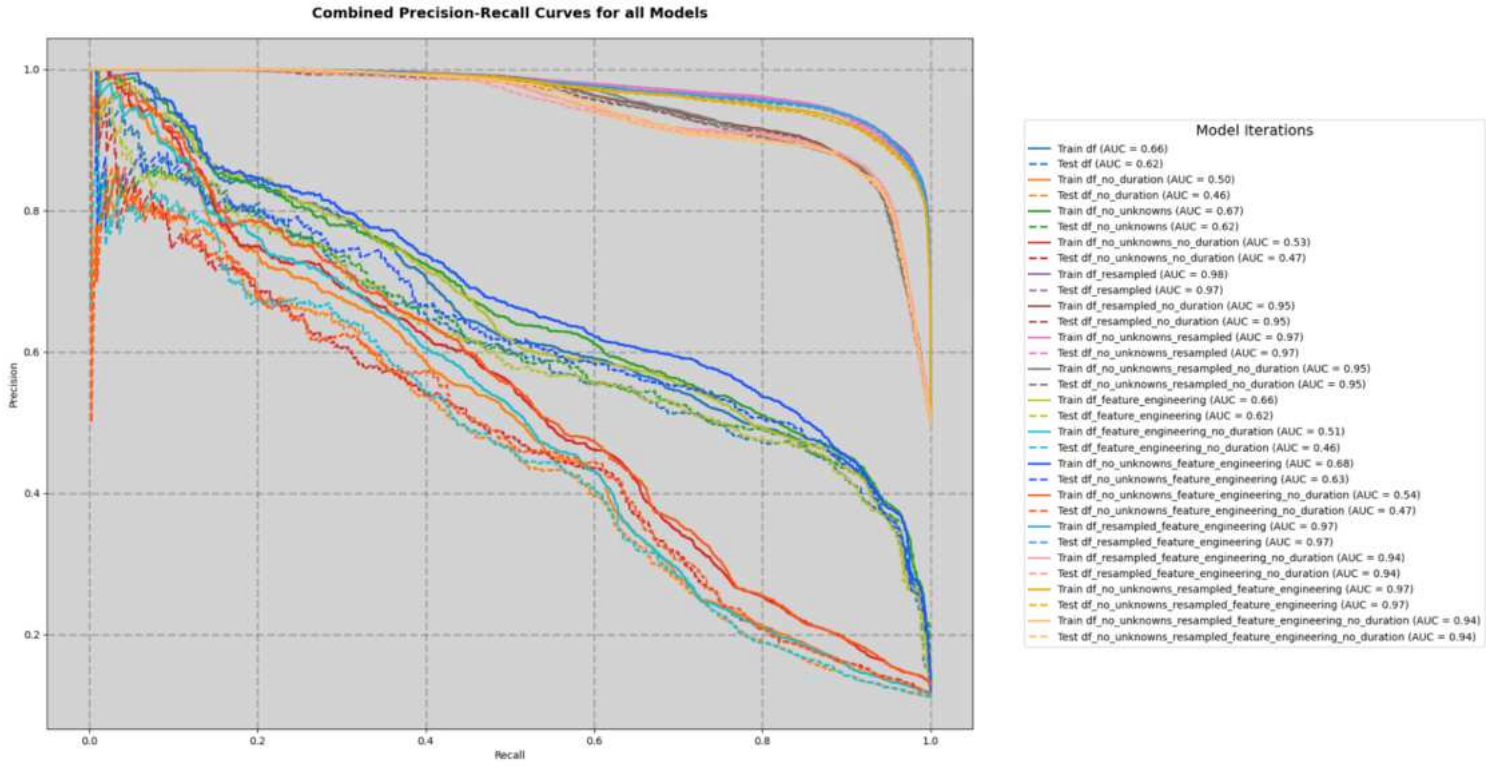

ROC Curve for df_no_unknowns_feature_engineering_no_duration
True Positive Rate
False Positive Rate
Train ROC curve (area = 1.00)
Test ROC curve (area = 0.78)
Confusion Matrix
Actual label
Predicted label
5161
213
490
231


ROC Curve for df_resampled_feature_engineering
True Positive Rate
False Positive Rate
Train ROC curve (area = 1.00)
Test ROC curve (area = 0.99)
Confusion Matrix
Actual label
Predicted label
7035
271
409
6898
ROC Curve for df_resampled_feature_engineering_no_duration
True Positive Rate
False Positive Rate
Train ROC curve (area = 1.00)
Test ROC curve (area = 0.97)
Confusion Matrix
Actual label
Predicted label
6995
311
590
6717
ROC Curve for df_no_unknowns_resampled_feature_engineering
True Positive Rate
False Positive Rate
Train ROC curve (area = 1.00)
Test ROC curve (area = 0.99)
Confusion Matrix
Actual label
Predicted label
5095
229
306
5016

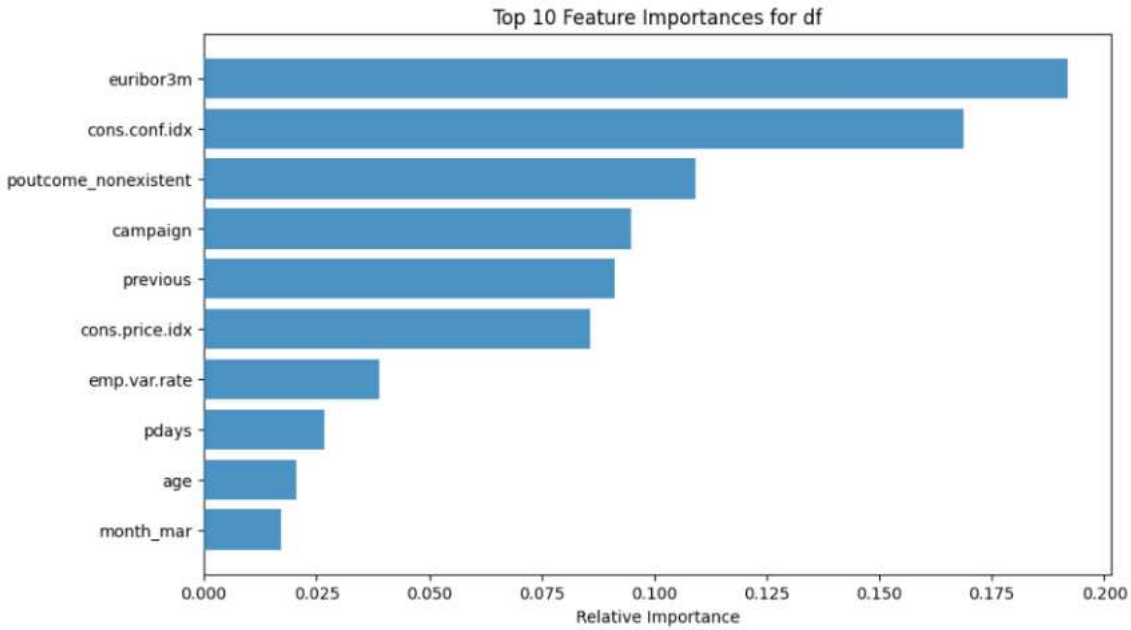


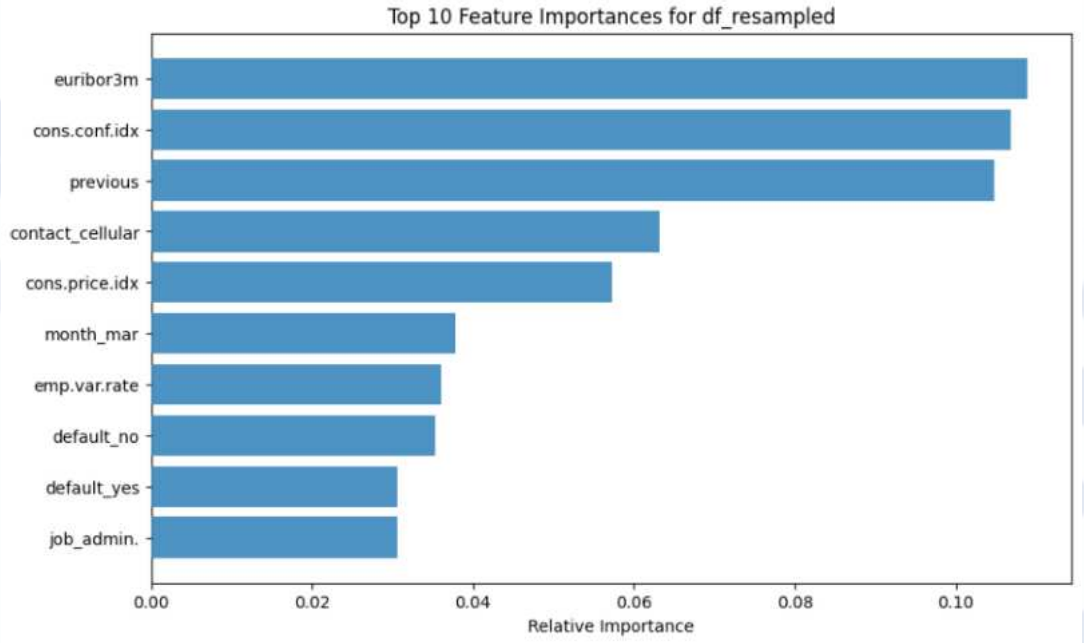
Ap. Figure 20 (set) Random Forest Classification Model iterations performed per data frame, with corresponding Confusion matrix.

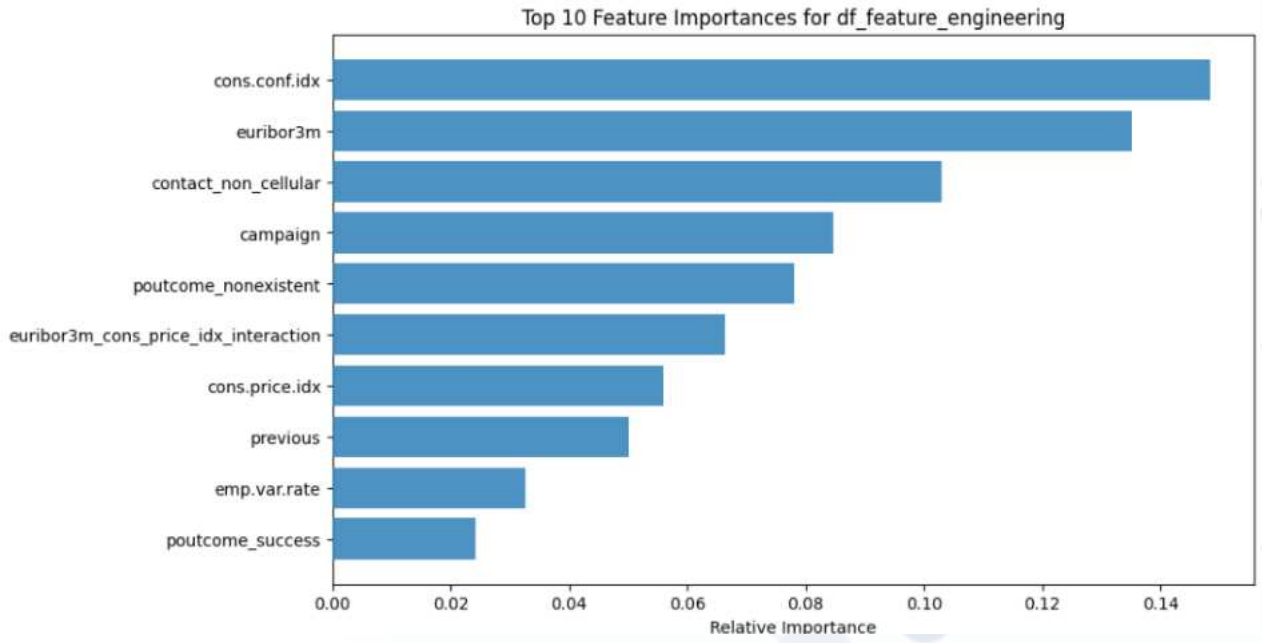




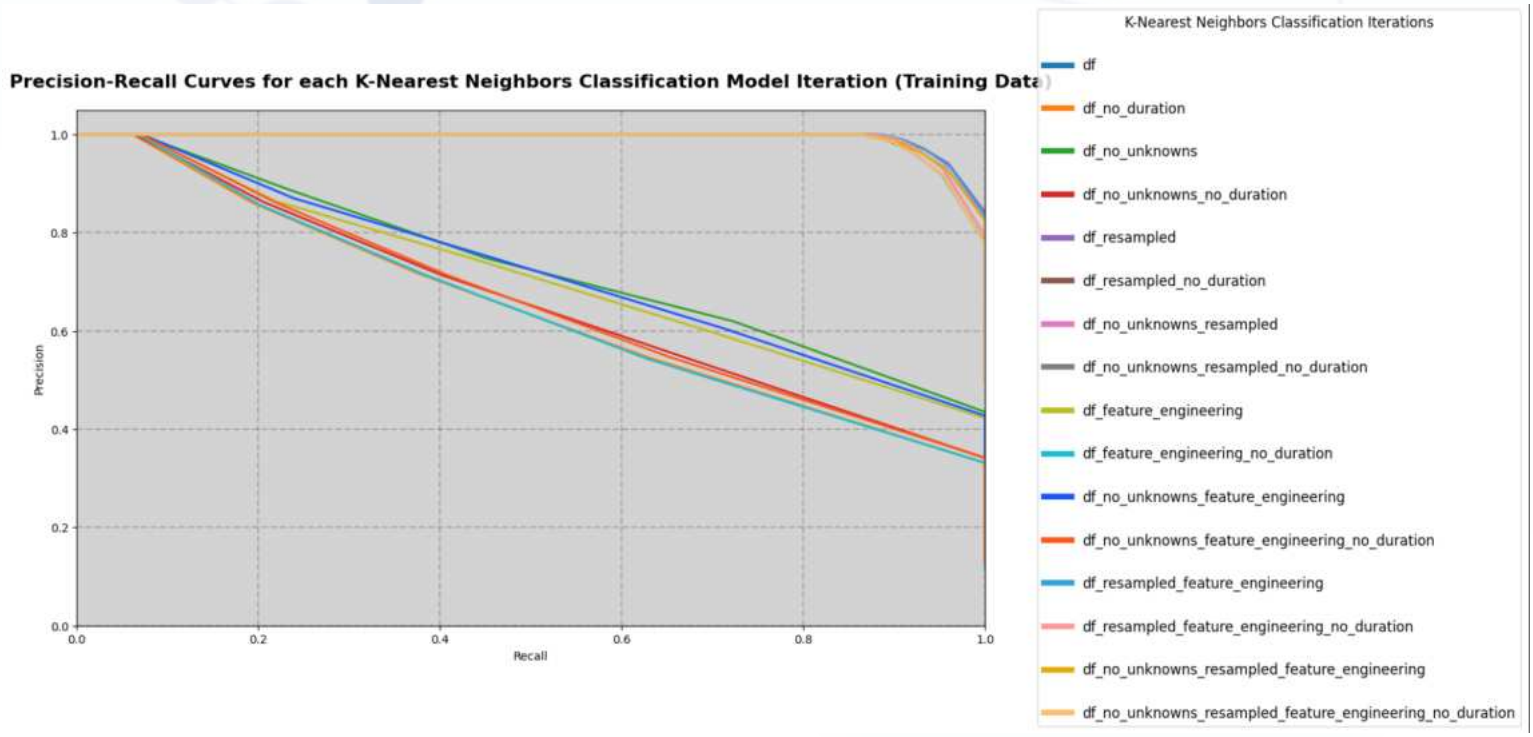
Ap. Figure 21 (set) Training and Testing Tuned Random Forest Classification Model iterations,
ROC curves and Precision-Recall curves –
(params:: n_estimators: 100, min_child_weight: 5-7, min_samples_split: 10-15, min_samples_leaf: 6-8, max_features: sqrt, log2)

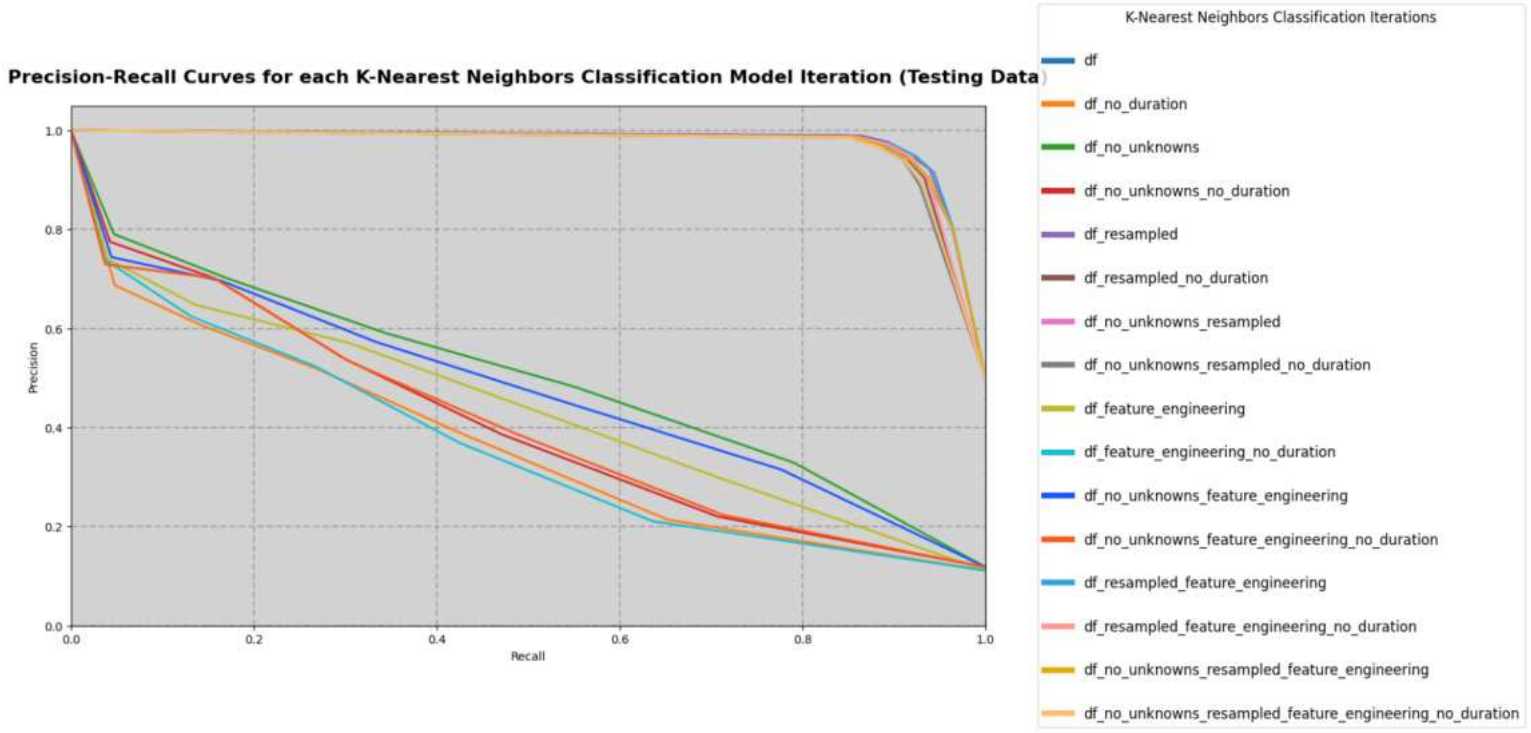






Ap. Figure 22 (set) Random Forest Feature importance(s) per key data frame.





Ap. Figure 23 (set) Training and Testing KNN Model iterations,
ROC curves and Precision-Recall curves