FINAL PROJECT PRESENTATION

Deepak Roy, Jerimi Blurton, Matt Eland, Matthew Louthan, Nicholas Norris

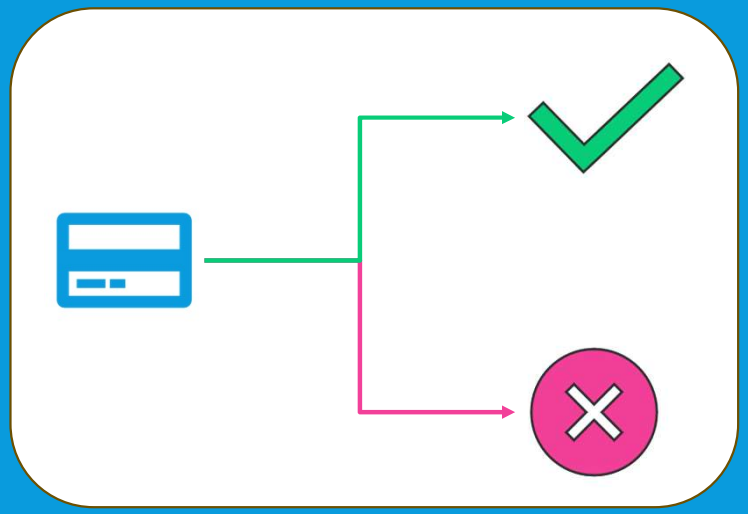
PART 1

Project Overview, Data Overview, and Data Exploration

PROBLEM DESCRIPTION & ML TYPE

- **Binary Classification (Supervised Learning) Problem:**

  Given known information on credit card history,
  predict if a card will default.



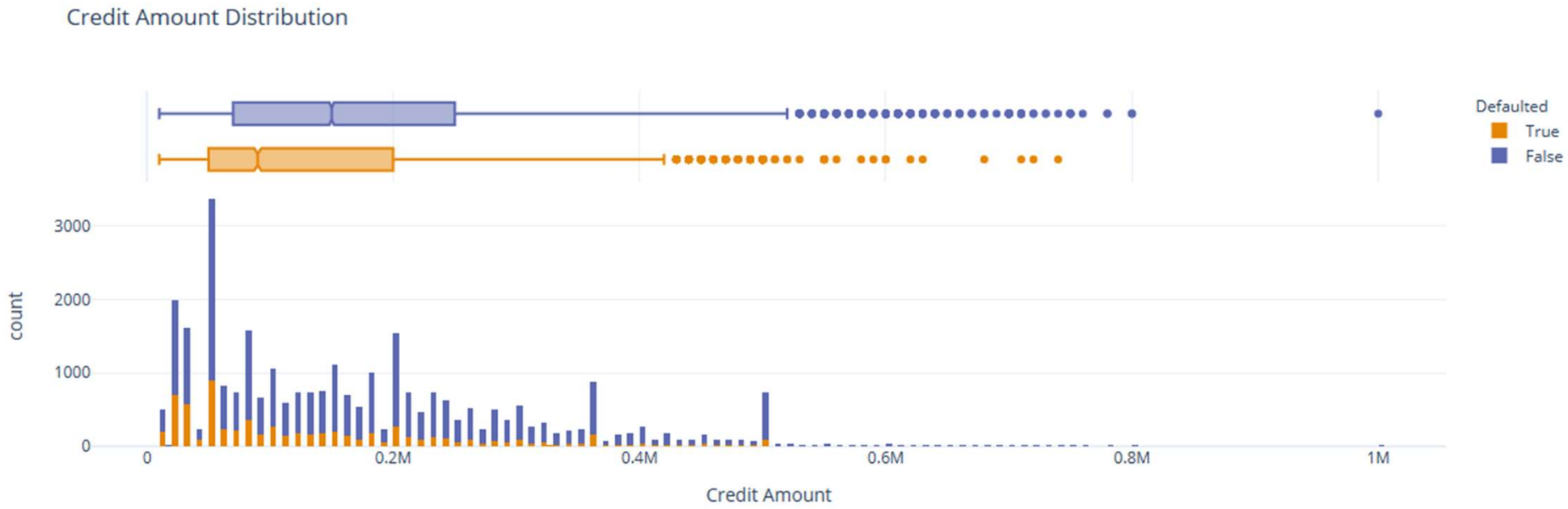
FEATURE AND TARGET VARIABLES

- **Features are X1-X23**
  - **X1 Credit amount**
  - **X2-5 –Gender, Education, Marital Status & Age**
  - **X6-11 – Repayment Delays Sep – Apr**
  - **X12-17 – Monthly bill Sep - Apr**
  - **X18-23 – Prior payment Sep – Apr**
- **Target Variable / Label**
  - **Y – Defaulted (1 or 0)**

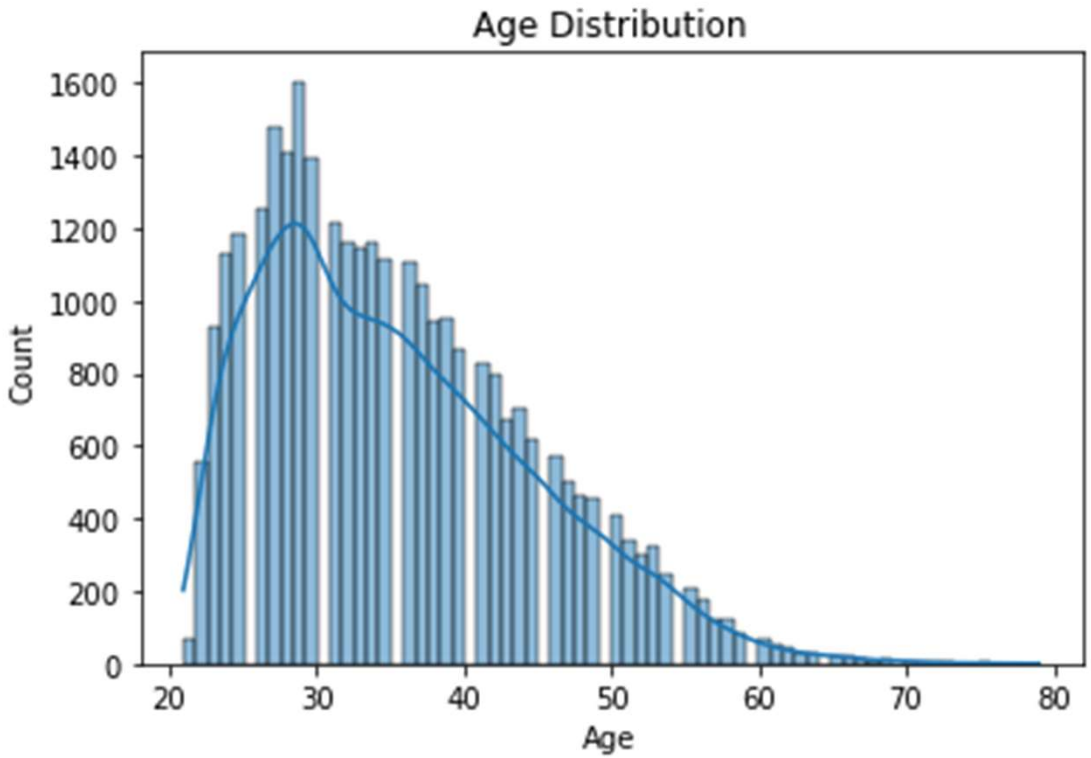
	Data Type	Missing Values
ID	int64	0
Credit Amount	int64	0
Gender	int64	0
Education	int64	0
Marital Status	int64	0
Age in Years	int64	0
Repay Delay Sep-2005	int64	0
Repay Delay Aug-2005	int64	0
Repay Delay Jul-2005	int64	0
Repay Delay Jun-2005	int64	0
Repay Delay May-2005	int64	0
Repay Delay Apr-2005	int64	0
Bill Sep-2005	int64	0
Bill Aug-2005	int64	0
Bill Jul-2005	int64	0
Bill Jun-2005	int64	0
Bill May-2005	int64	0
Bill Apr-2005	int64	0
Prior Pay Sep-2005	int64	0
Prior Pay Aug-2005	int64	0
Prior Pay Jul-2005	int64	0
Prior Pay Jun-2005	int64	0
Prior Pay May-2005	int64	0
Prior Pay Apr-2005	int64	0
Defaulted	int64	0

DATA EXPLORATION PROCESS

	ID	Credit Amount	Gender	Education	Marital Status	Age in Years	Repay Delay Sep-2005	Repay Delay Aug-2005	Repay Delay Jul-2005	Repay Delay Jun-2005
count	30000.000000	30000.000000	30000.000000	30000.000000	30000.000000	30000.000000	30000.000000	30000.000000	30000.000000	30000.000000
mean	15000.500000	167484.322667	1.603733	1.853133	1.551867	35.485500	-0.016700	-0.133767	-0.166200	-0.220667
std	8660.398374	129747.661567	0.489129	0.790349	0.521970	9.217904	1.123802	1.197186	1.196868	1.169139
min	1.000000	10000.000000	1.000000	0.000000	0.000000	21.000000	-2.000000	-2.000000	-2.000000	-2.000000
25%	7500.750000	50000.000000	1.000000	1.000000	1.000000	28.000000	-1.000000	-1.000000	-1.000000	-1.000000
50%	15000.500000	140000.000000	2.000000	2.000000	2.000000	34.000000	0.000000	0.000000	0.000000	0.000000
75%	22500.250000	240000.000000	2.000000	2.000000	2.000000	41.000000	0.000000	0.000000	0.000000	0.000000
max	30000.000000	1000000.000000	2.000000	6.000000	3.000000	79.000000	8.000000	8.000000	8.000000	8.000000



DATA VISUALIZATION – CREDIT AMOUNT

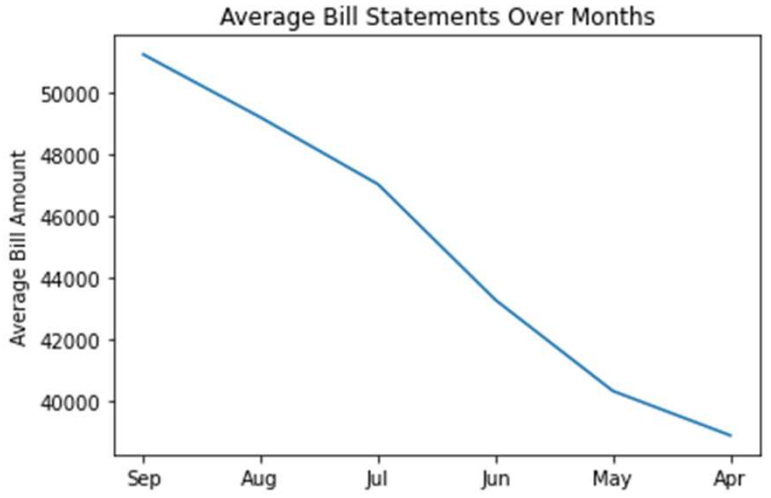


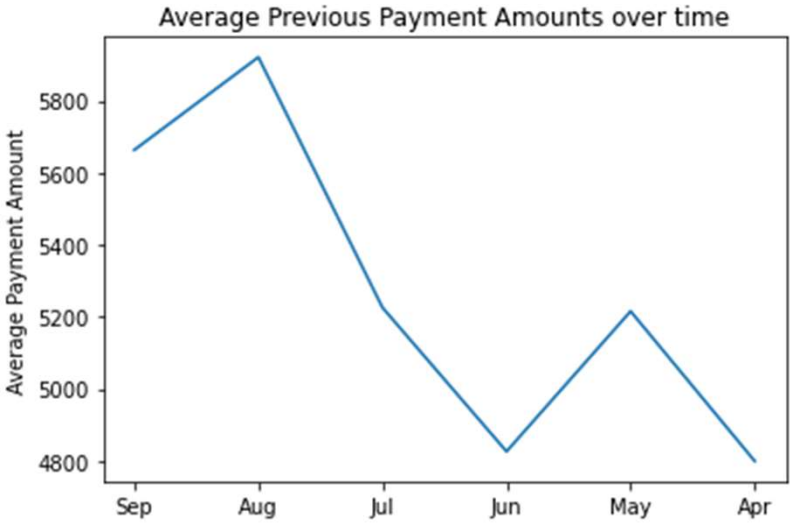
AGE HISTOGRAM

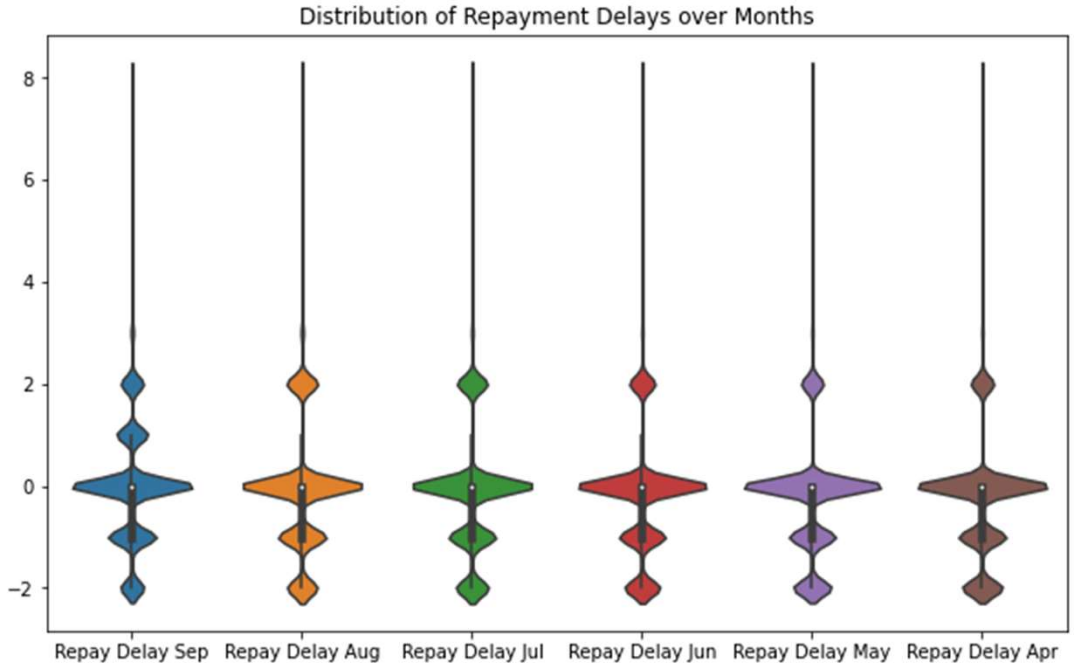
BILLING & PAY TRENDS

The average credit card bill declines nearly linearly

Recent months tend to have larger payments on average



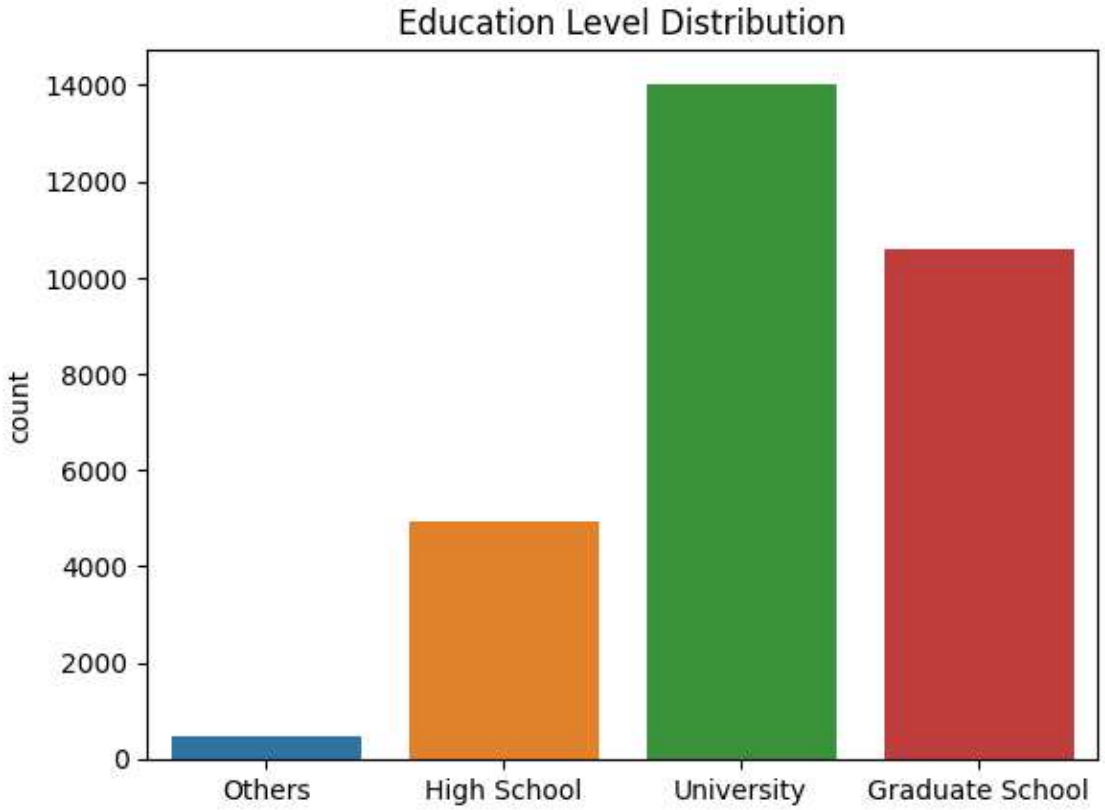




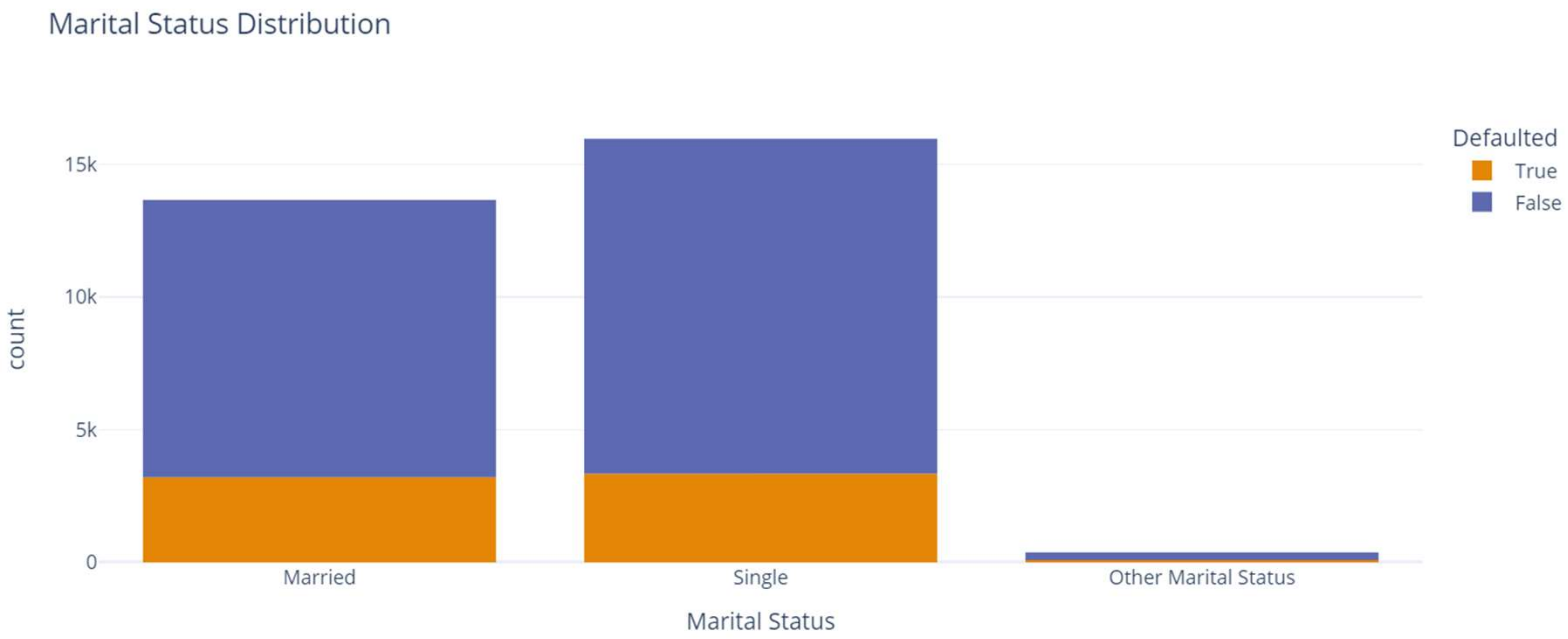
REPAYMENT DELAYS (IN MONTHS)

0 or -1 Represent on-time payments
-2 Assumed to be paid early

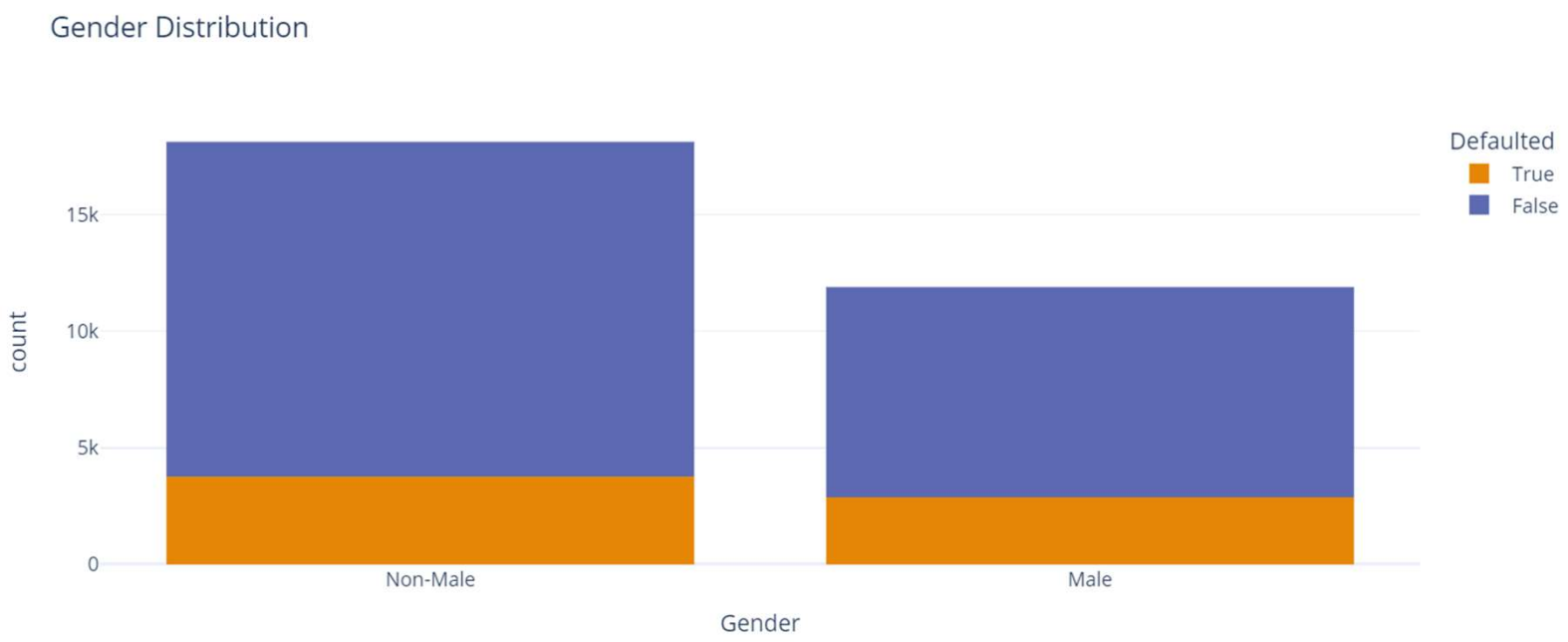
All negative values were later set to zero as part of data cleaning


Education Level Distribution
14000
12000
10000
8000
6000
4000
2000
0
count
Others
High School
University
Graduate School

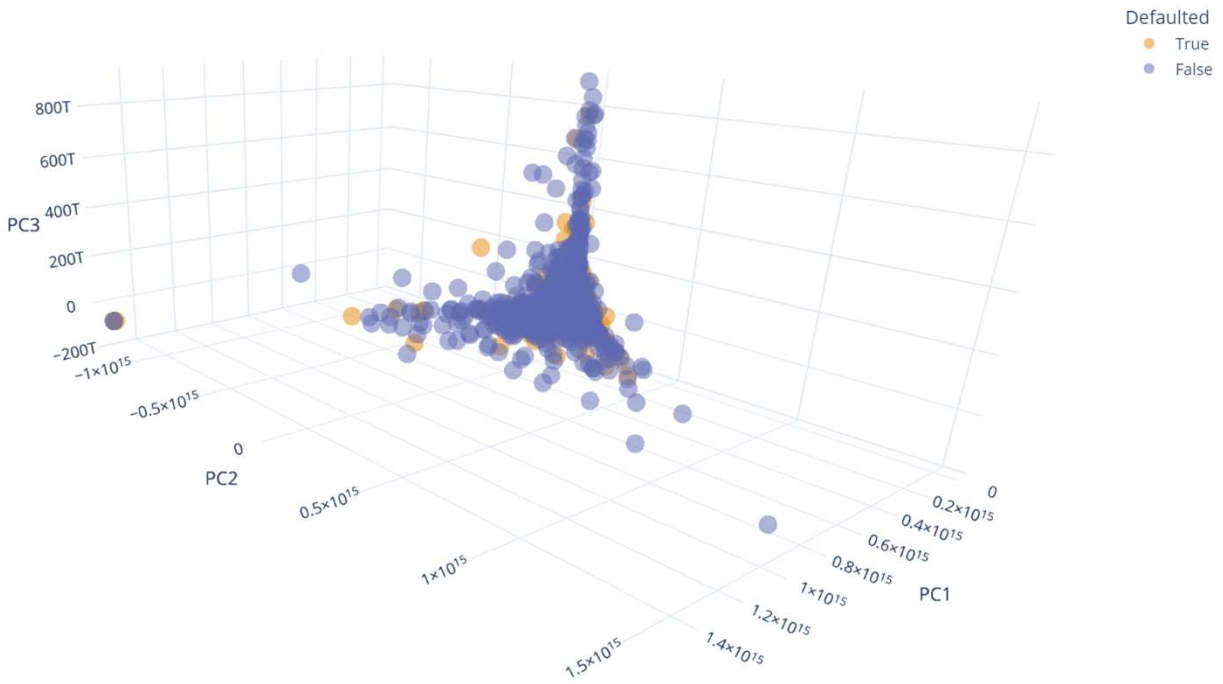
EDUCATION



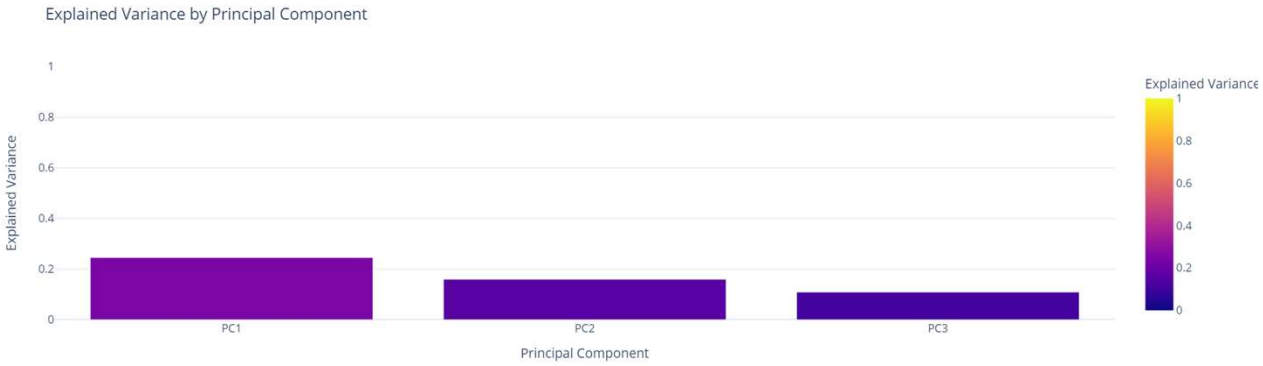
MARITAL STATUS


Gender Distribution
Defaulted
True
False
15k
10k
5k
0
count
Non-Male
Male
Gender

GENDER



PRINCIPAL COMPONENT ANALYSIS

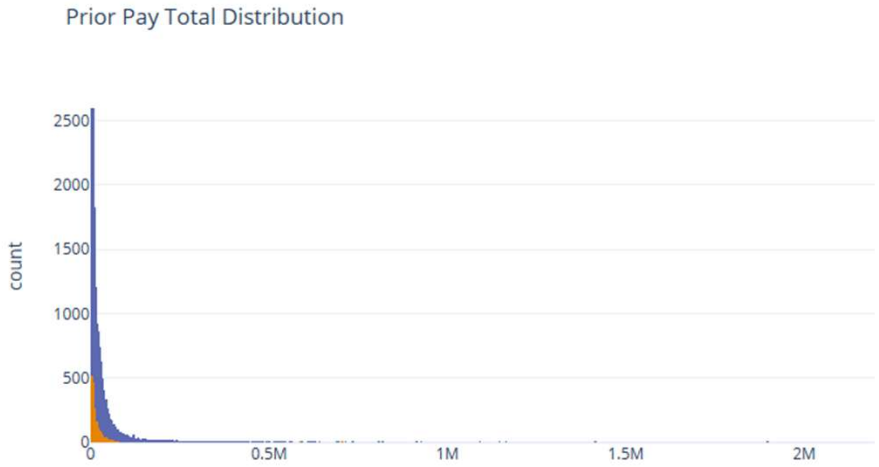


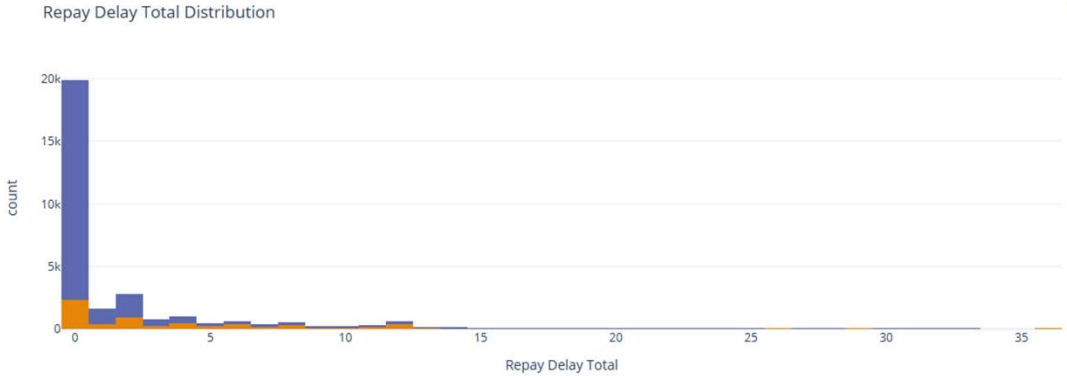
PART 2

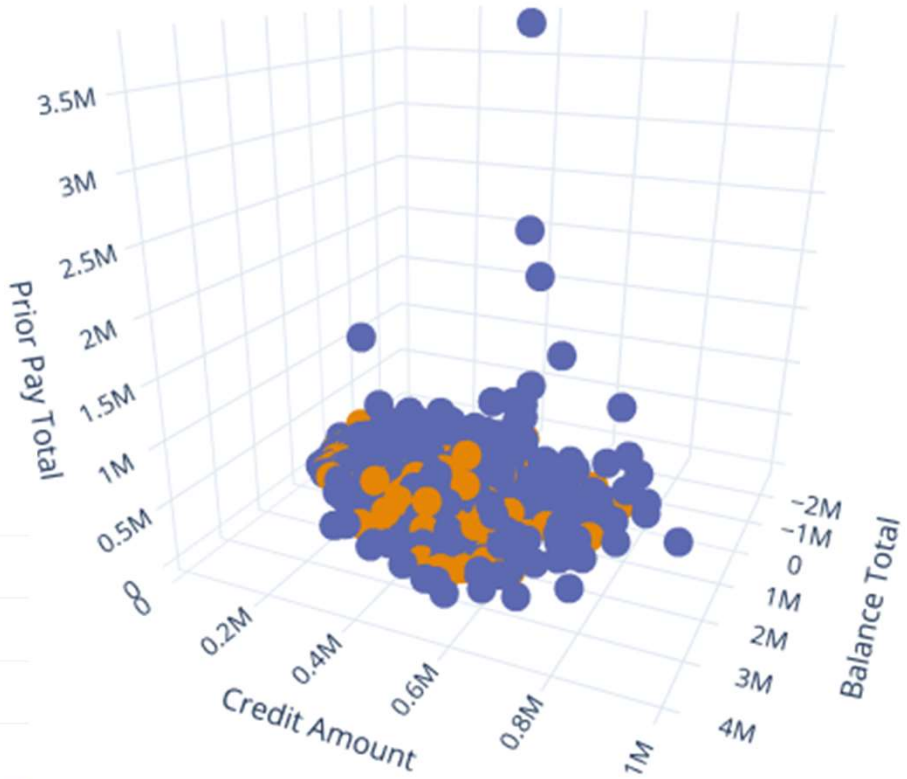
Data Wrangling, Model Training, & Model Evaluation

DATA CLEANING

- Used the ID column as the row index
- Replaced negative repay delays with 0
- Replaced Gender with Is Male
- Replaced Marital Status with Is Married
- One-hot encoded Education
- Removed Outliers





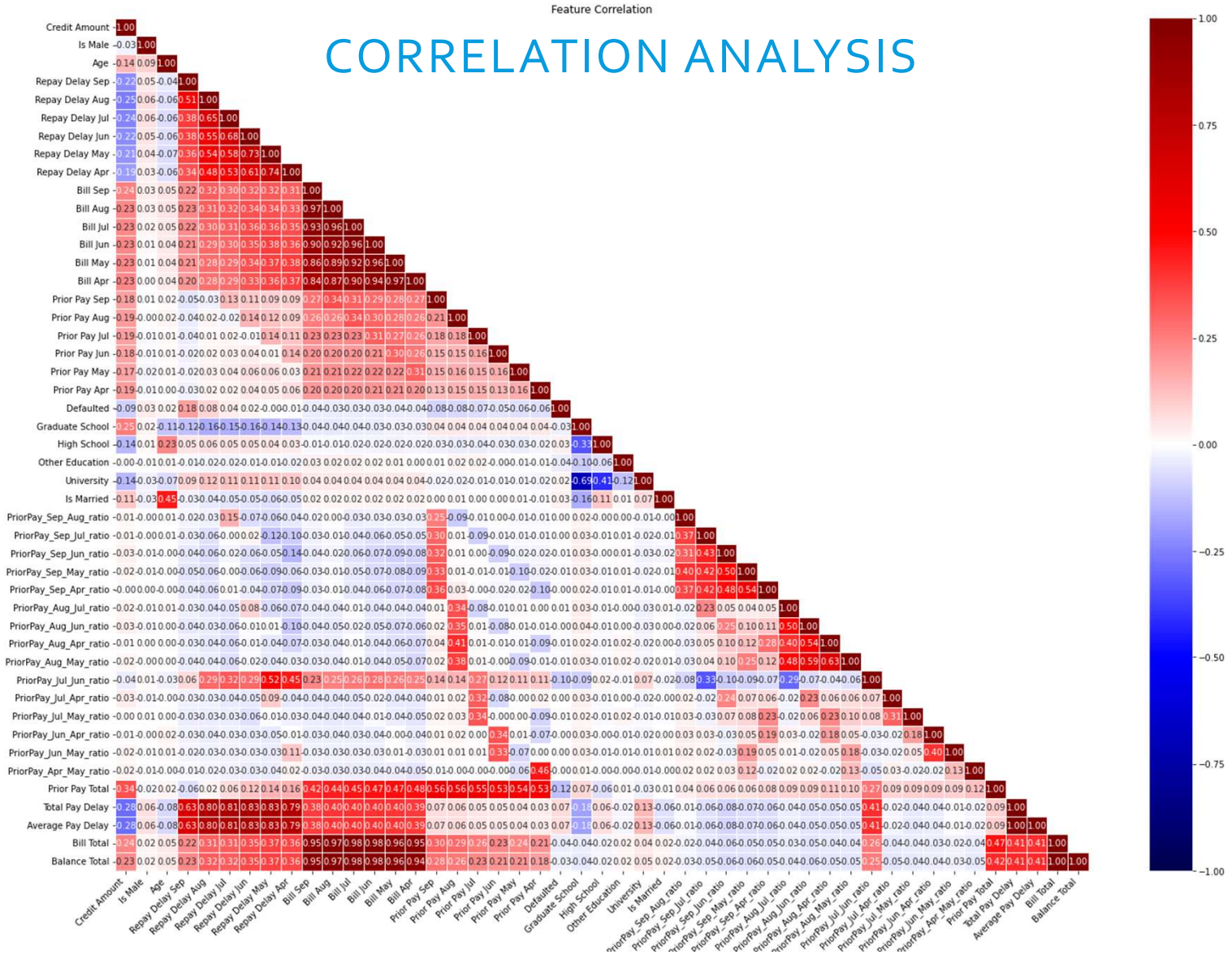


OUTLIER IDENTIFICATION

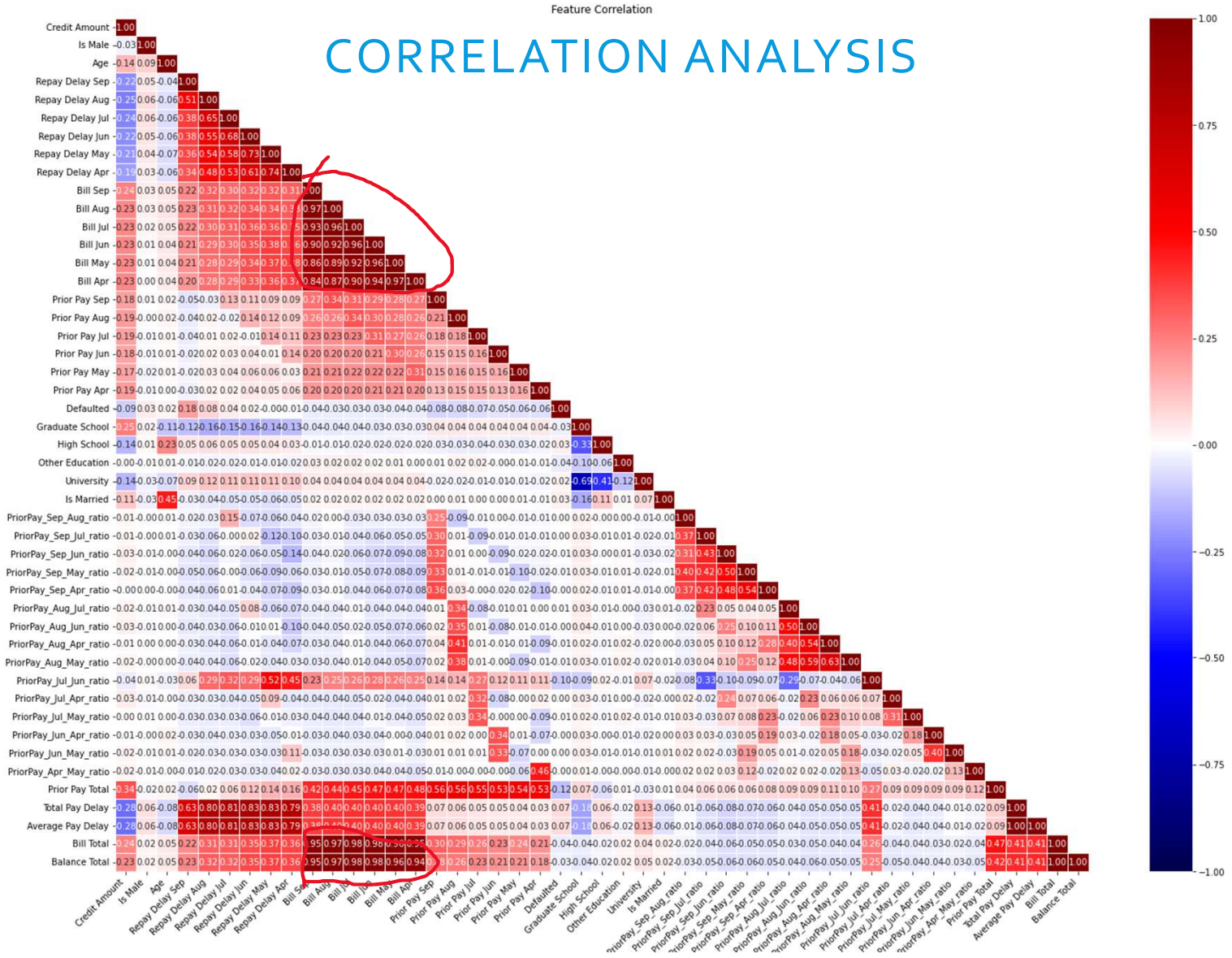
FEATURE ENGINEERING

- Added balance columns (bill – prior pay)
- Added total and average columns for
  - Prior payment
  - Repay delay
  - Bill
- Added ratio columns for repay delay differences between months

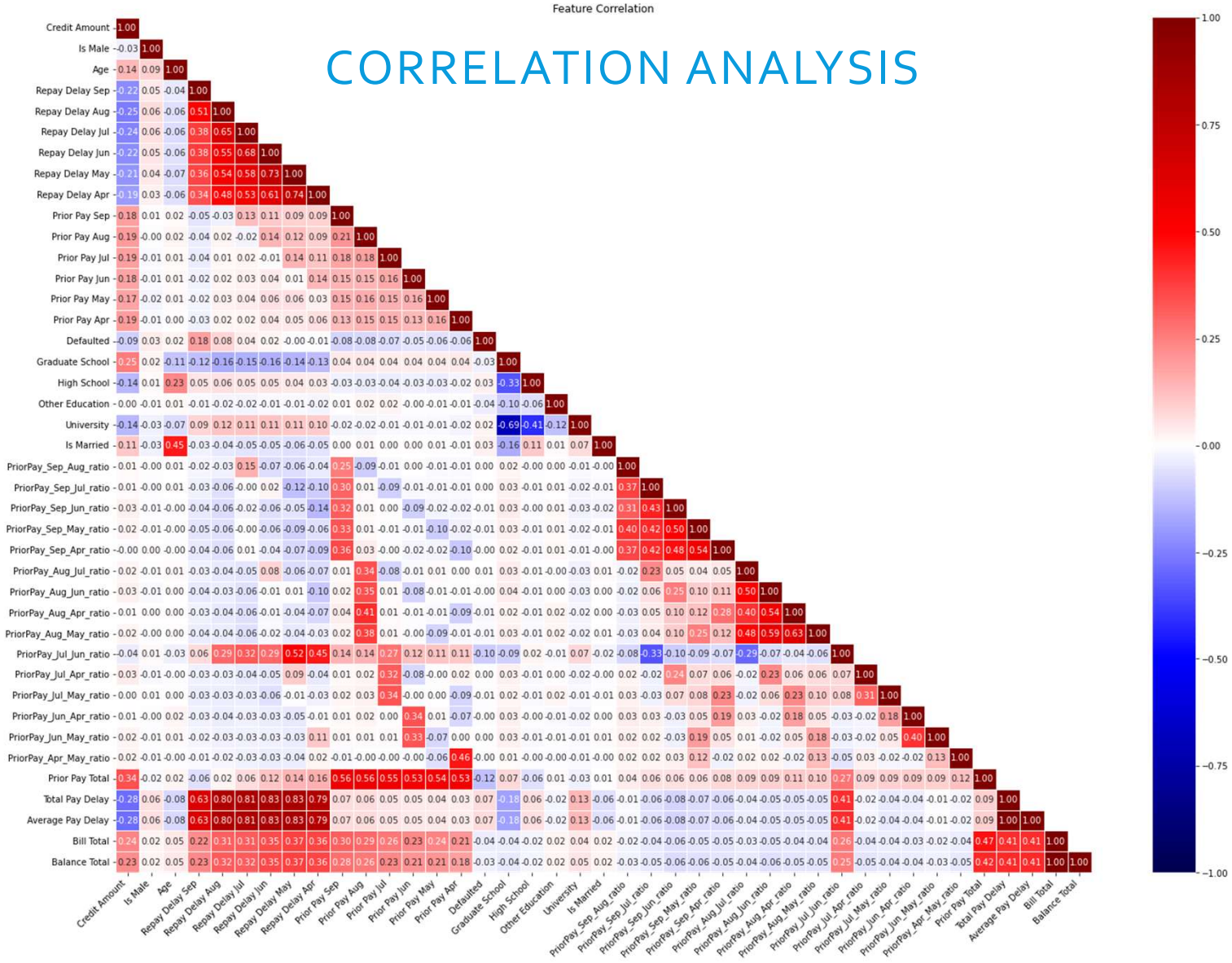
CORRELATION ANALYSIS



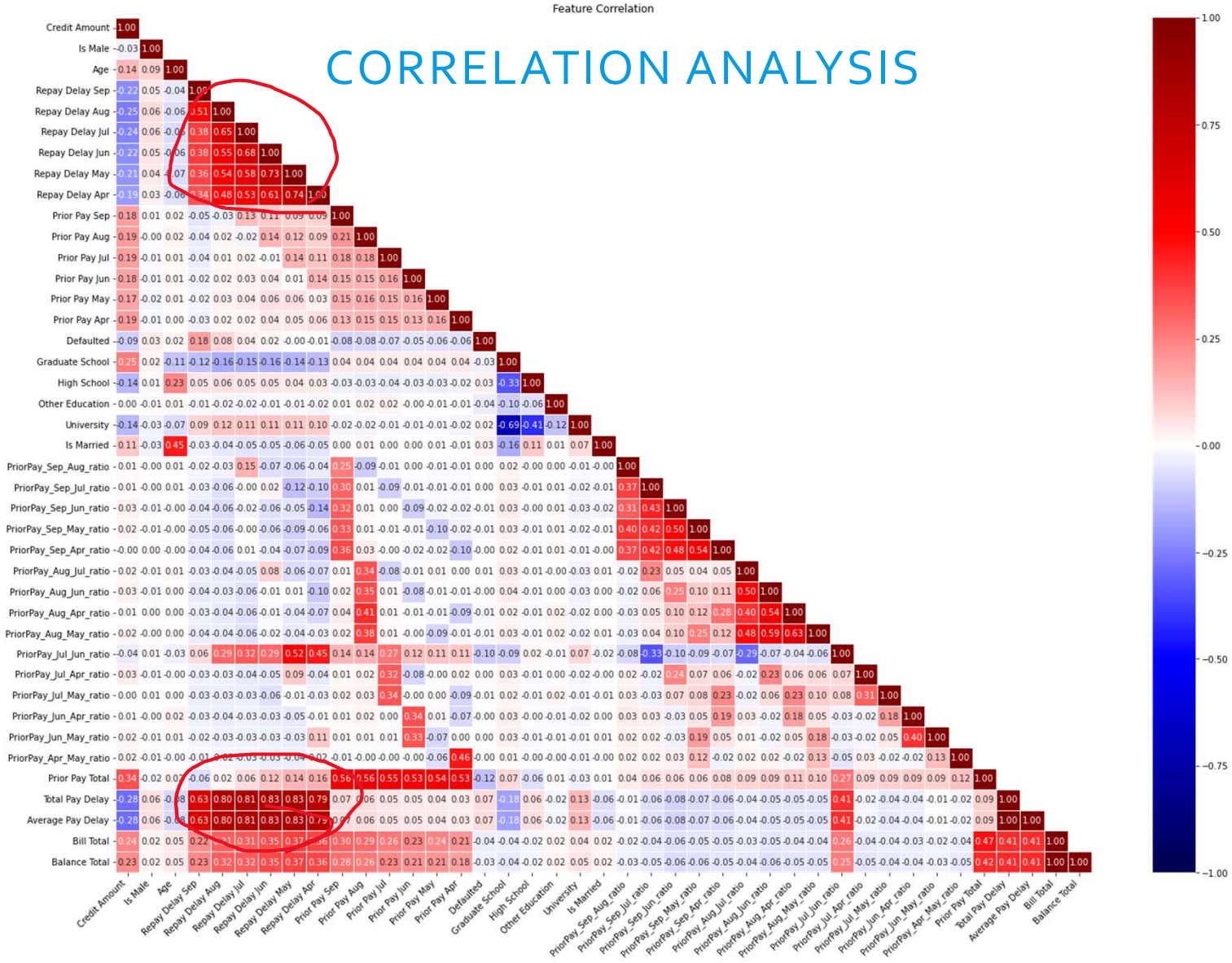
CORRELATION ANALYSIS



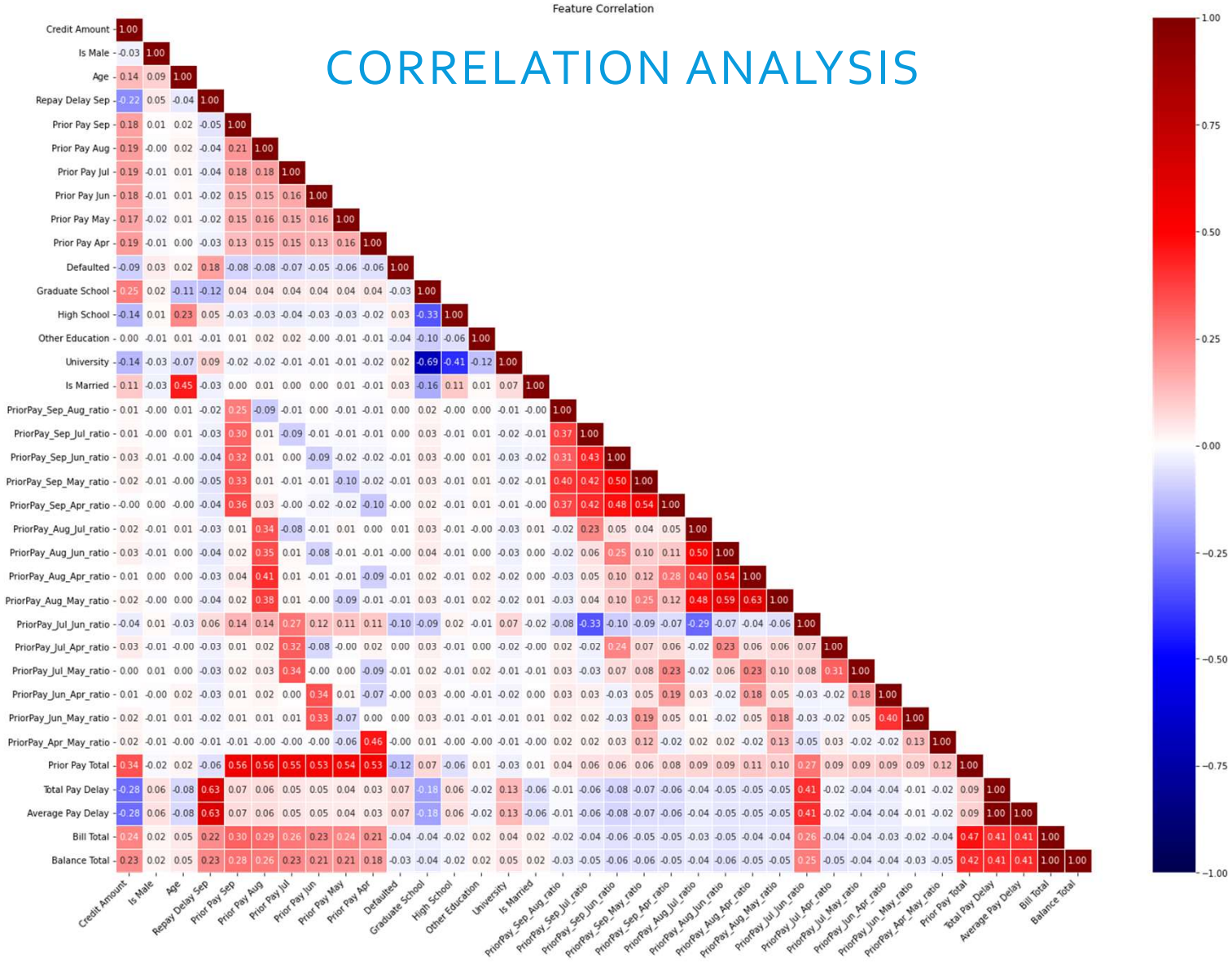
CORRELATION ANALYSIS



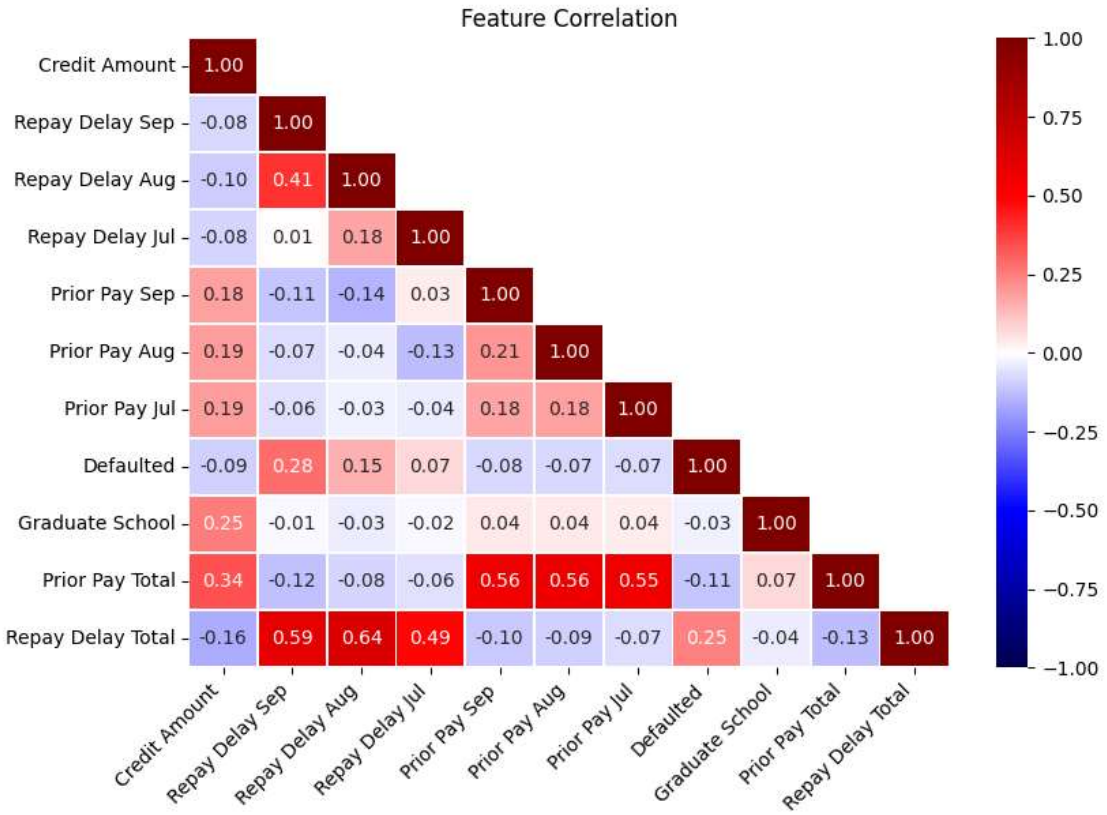
CORRELATION ANALYSIS

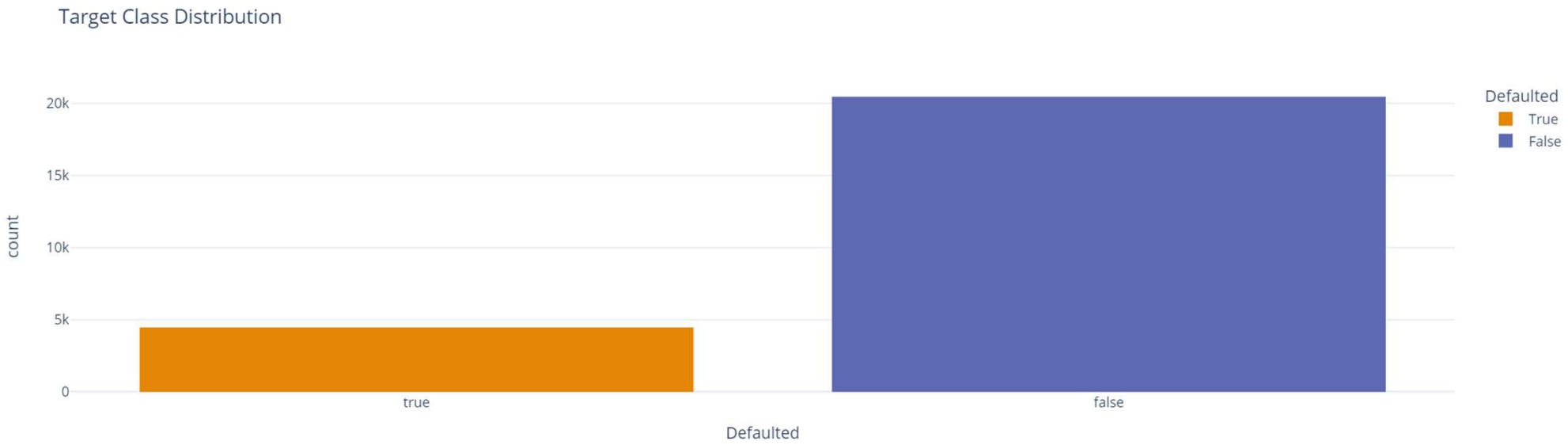


CORRELATION ANALYSIS

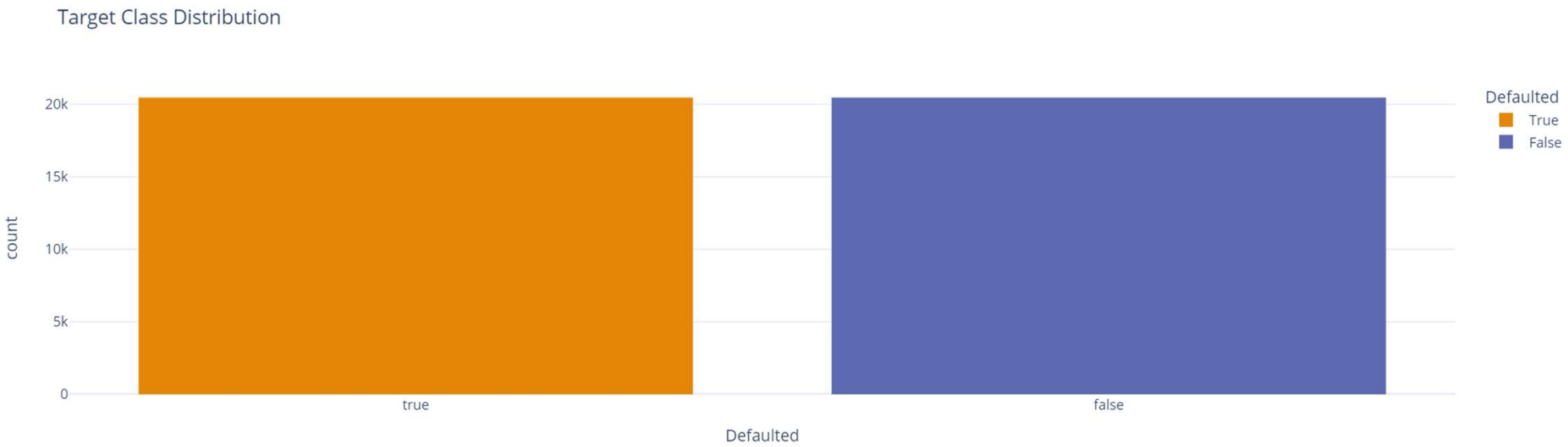


CORRELATION ANALYSIS





CLASS IMBALANCE



SMOTE APPLIED

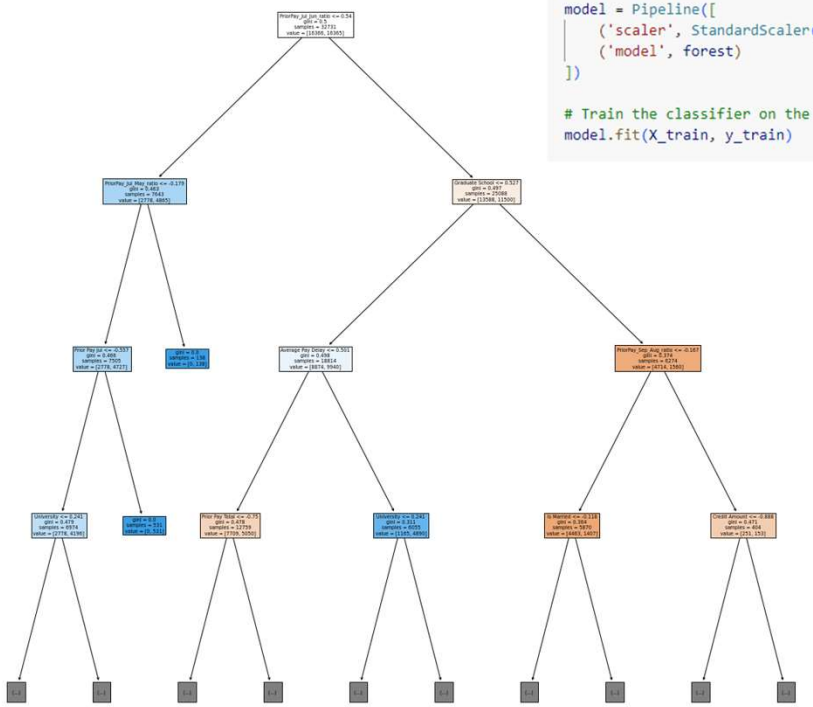
TEST / TRAIN SPLIT

- 20 % Test, 80 % Train
  (prescribed by assignment)
- Stratified by
  - Defaulted
  - Is Male
  - Age // 10
- Removed Bias Columns
  - Is Male
  - Age

```
Defaulted_True_Male_True_Age_3      5531
Defaulted_False_Male_False_Age_3    4794
Defaulted_False_Male_False_Age_2    4416
Defaulted_True_Male_False_Age_3     3236
Defaulted_True_Male_True_Age_2      3176
Defaulted_False_Male_True_Age_3     2990
Defaulted_True_Male_True_Age_4      2926
Defaulted_True_Male_False_Age_2     2631
Defaulted_False_Male_False_Age_4    2532
Defaulted_False_Male_True_Age_2     2174
Defaulted_False_Male_True_Age_4     1795
Defaulted_True_Male_False_Age_4     1650
Defaulted_False_Male_False_Age_5     836
Defaulted_True_Male_True_Age_5       734
Defaulted_False_Male_True_Age_5      707
Defaulted_True_Male_False_Age_5      405
Defaulted_True_Male_True_Age_6       103
Defaulted_False_Male_True_Age_6      100
Defaulted_False_Male_False_Age_6      95
Defaulted_True_Male_False_Age_6       58
Defaulted_False_Male_False_Age_7       9
Defaulted_False_Male_True_Age_7        9
Defaulted_True_Male_True_Age_7         4
Defaulted_True_Male_False_Age_7        3
Name: Stratify, dtype: int64
```

MODEL SELECTION

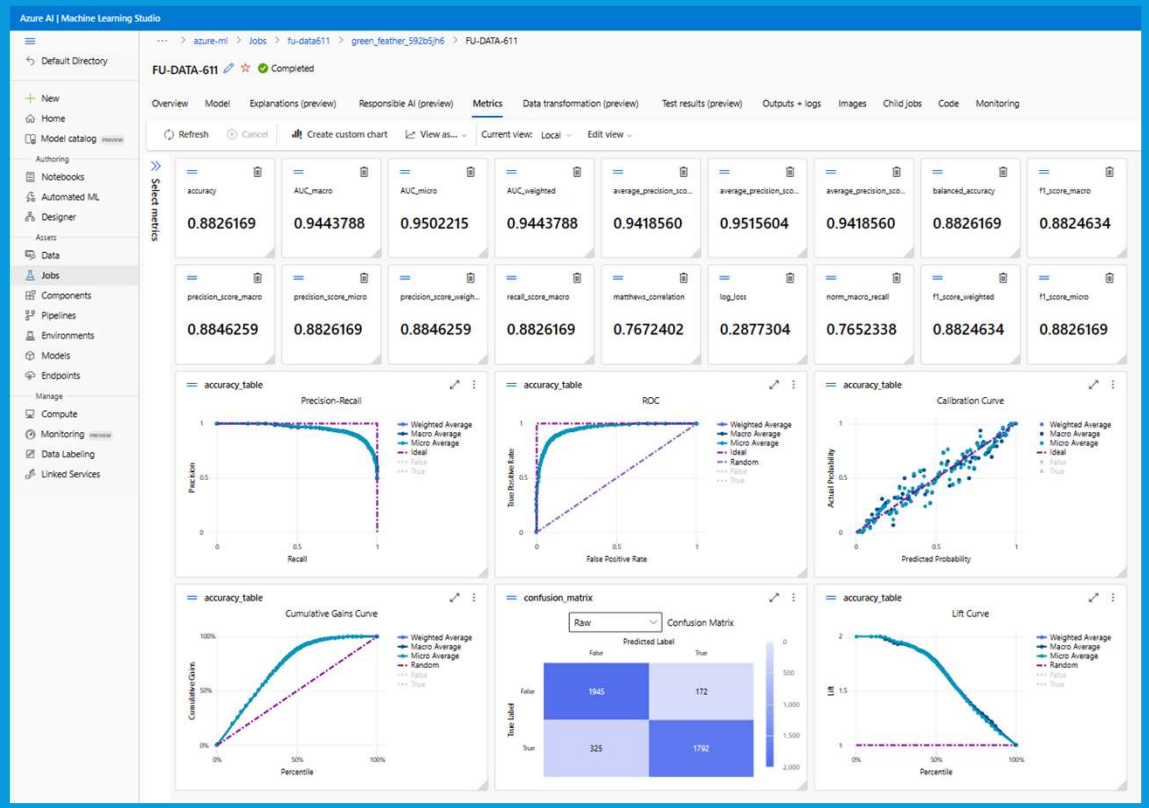
- Pipeline with Standard Scalar
- Models Considered
  - Logistic Regression
  - Decision Tree
  - **Random Forest**
  - K-Neighbors
  - Gradient Boosting
  - Ensemble Models
- Tuning
  - Grid Search / Manual Tuning
- Evaluation
  - Standard Metrics
  - Cross Validation
  - Explainable Models



```python
from sklearn.ensemble import RandomForestClassifier

# Create a pipeline with a standard scalar and the random forest classifier
from sklearn.pipeline import Pipeline

num_cores = 8

# Initialize the classifier with the best parameters
forest = RandomForestClassifier(
    bootstrap=False,
    max_depth=6,
    max_features='sqrt',
    min_samples_leaf=1,
    min_samples_split=2,
    n_estimators=150,
    random_state=123,
    n_jobs=num_cores
)

model = Pipeline([
    ('scaler', StandardScaler()), # Not really needed with a Random Forest
    ('model', forest)
])

# Train the classifier on the resampled training data
model.fit(X_train, y_train)
```

AUTOMATED ML

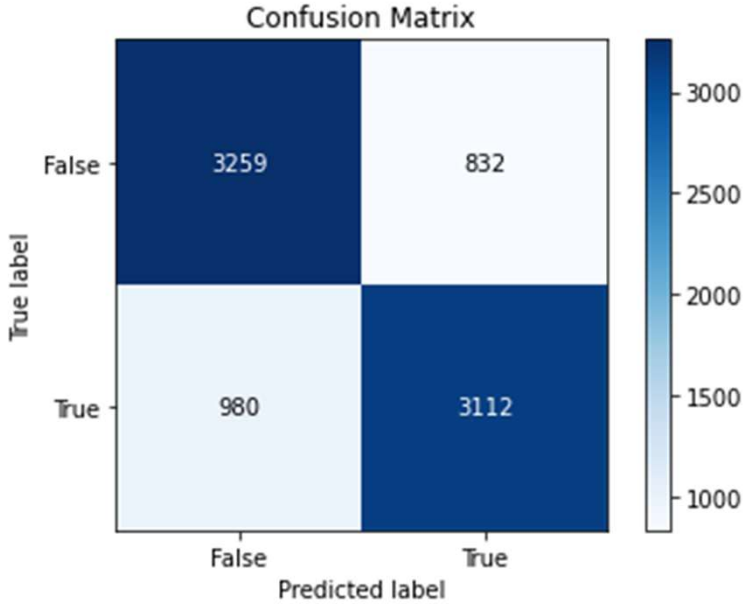
AZURE ML & AUTO-SKLEARN



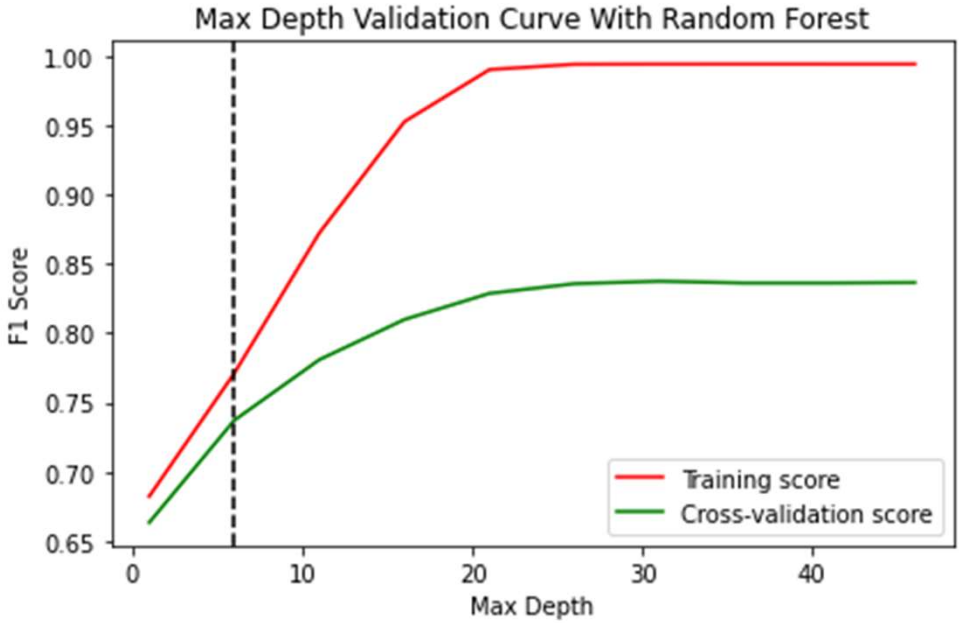
model_id	rank	ensemble_weight	type	cost	duration
312	1	0.04	gradient_boosting	0.188775	37.328512
176	2	0.06	gradient_boosting	0.190801	33.657909
469	3	0.04	gradient_boosting	0.190929	43.874617
751	4	0.06	gradient_boosting	0.191211	29.784710
684	5	0.02	gradient_boosting	0.191227	48.822893
736	6	0.08	gradient_boosting	0.191533	39.896581
692	7	0.02	gradient_boosting	0.191609	37.488533
666	8	0.14	gradient_boosting	0.192243	34.652723
695	9	0.02	gradient_boosting	0.192250	37.027887
661	10	0.06	gradient_boosting	0.192369	37.656005
722	11	0.02	gradient_boosting	0.192404	34.724431
229	12	0.02	gradient_boosting	0.192655	52.113459
639	13	0.02	gradient_boosting	0.192756	34.945801
493	14	0.02	gradient_boosting	0.192986	39.809637
311	15	0.08	gradient_boosting	0.193254	101.888010
191	16	0.14	gradient_boosting	0.193317	103.276103
353	17	0.02	gradient_boosting	0.193332	38.953808
443	18	0.14	gradient_boosting	0.193849	86.005936

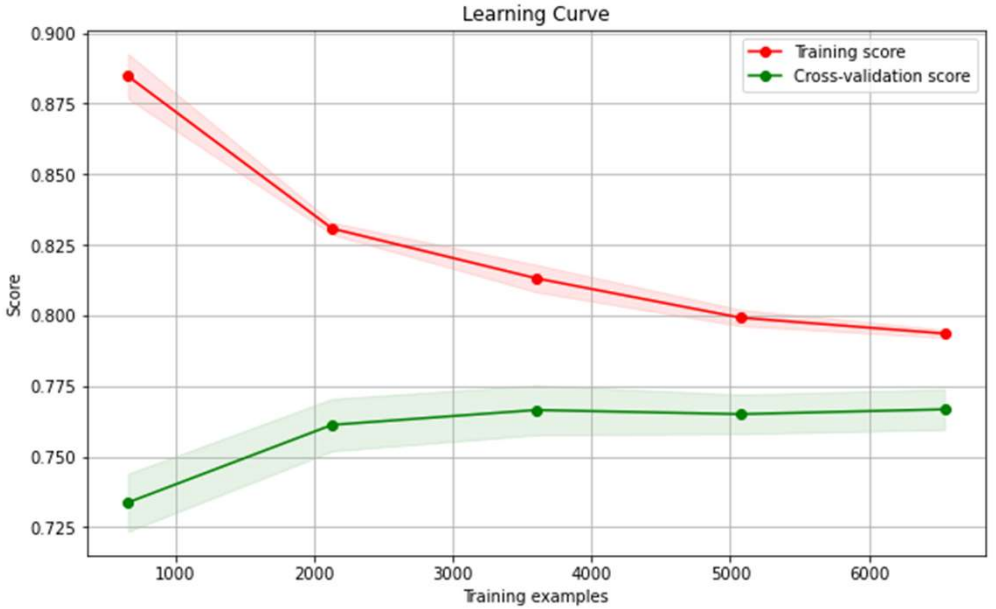
	precision	recall	f1-score	support
Not Defaulted	0.77	0.80	0.78	4091
Defaulted	0.79	0.76	0.77	4092
accuracy			0.78	8183
macro avg	0.78	0.78	0.78	8183
weighted avg	0.78	0.78	0.78	8183

F1 Score: 0.7745



MODEL EVALUATION METRICS

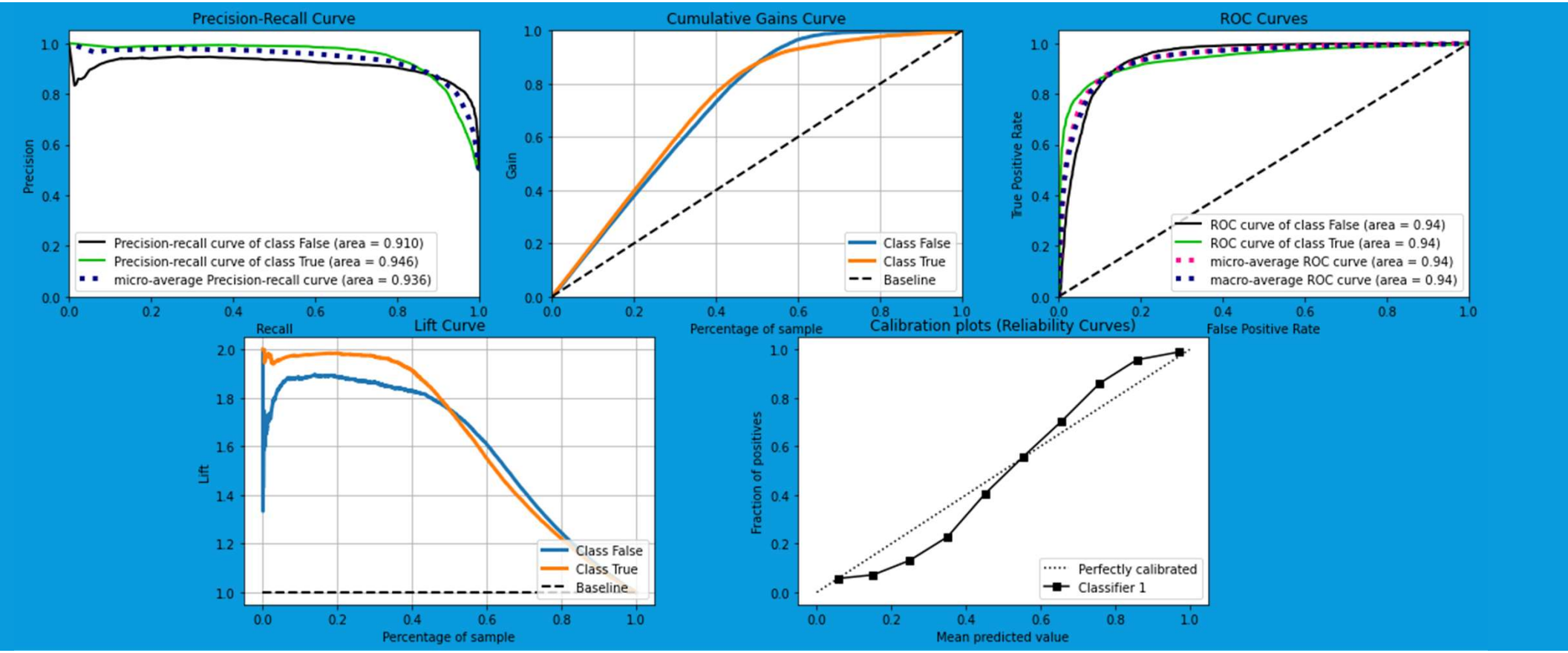


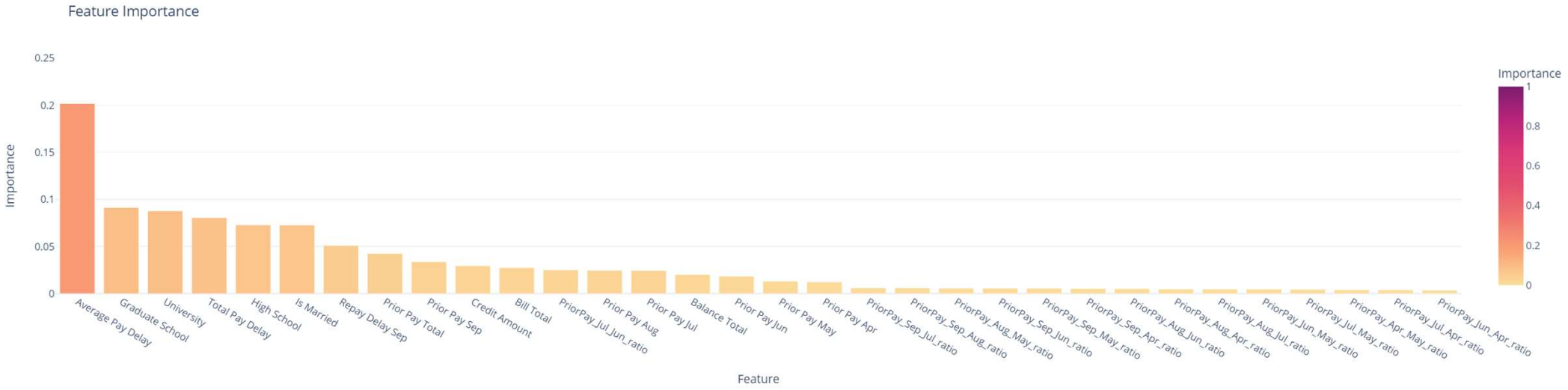


MODEL FITTING

Accuracy: 0.7671
Precision: 0.7783
Recall: 0.7381
F1: 0.7532
AUC: 0.8464

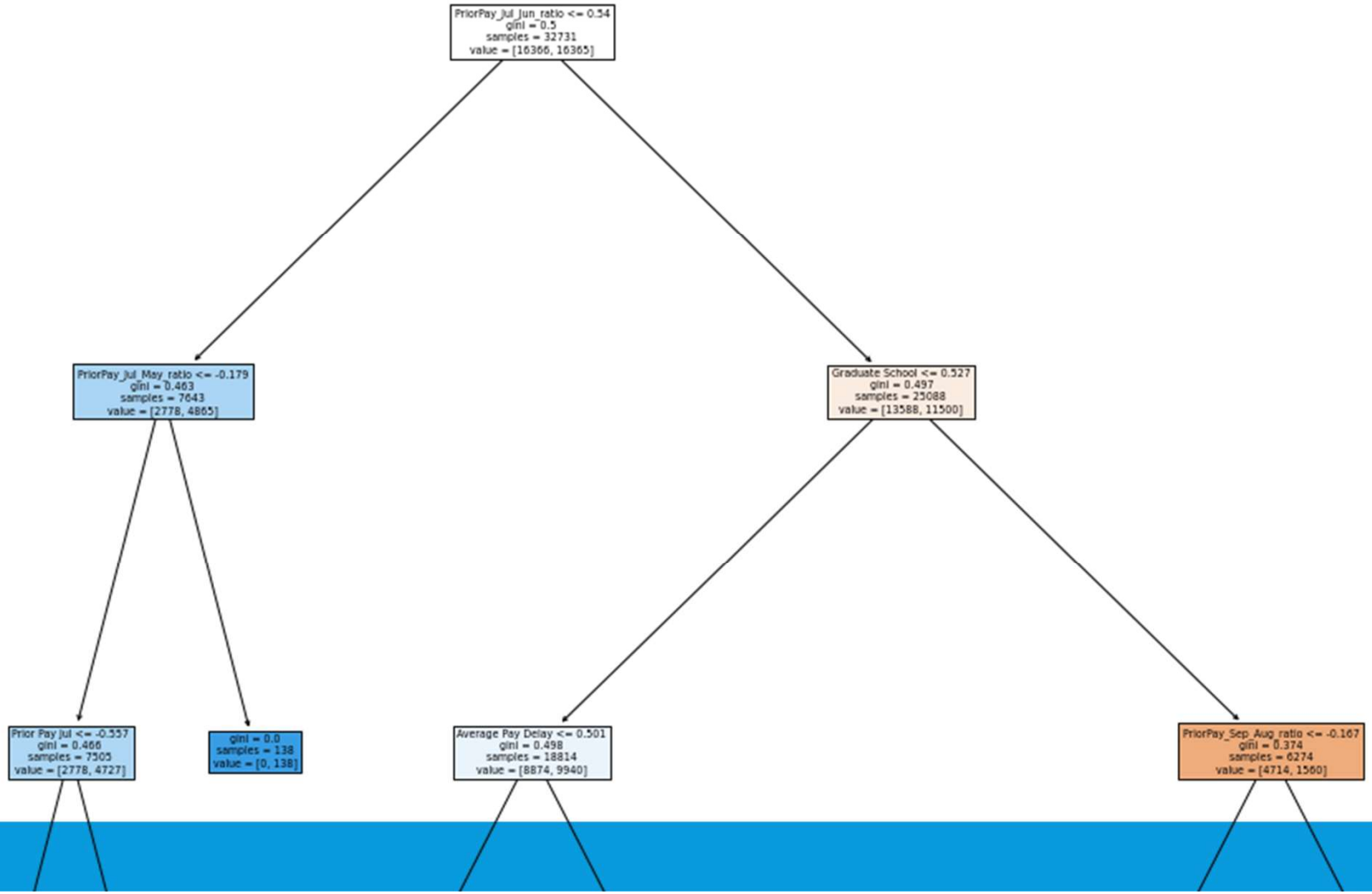
MODEL EVALUATION GRAPHS

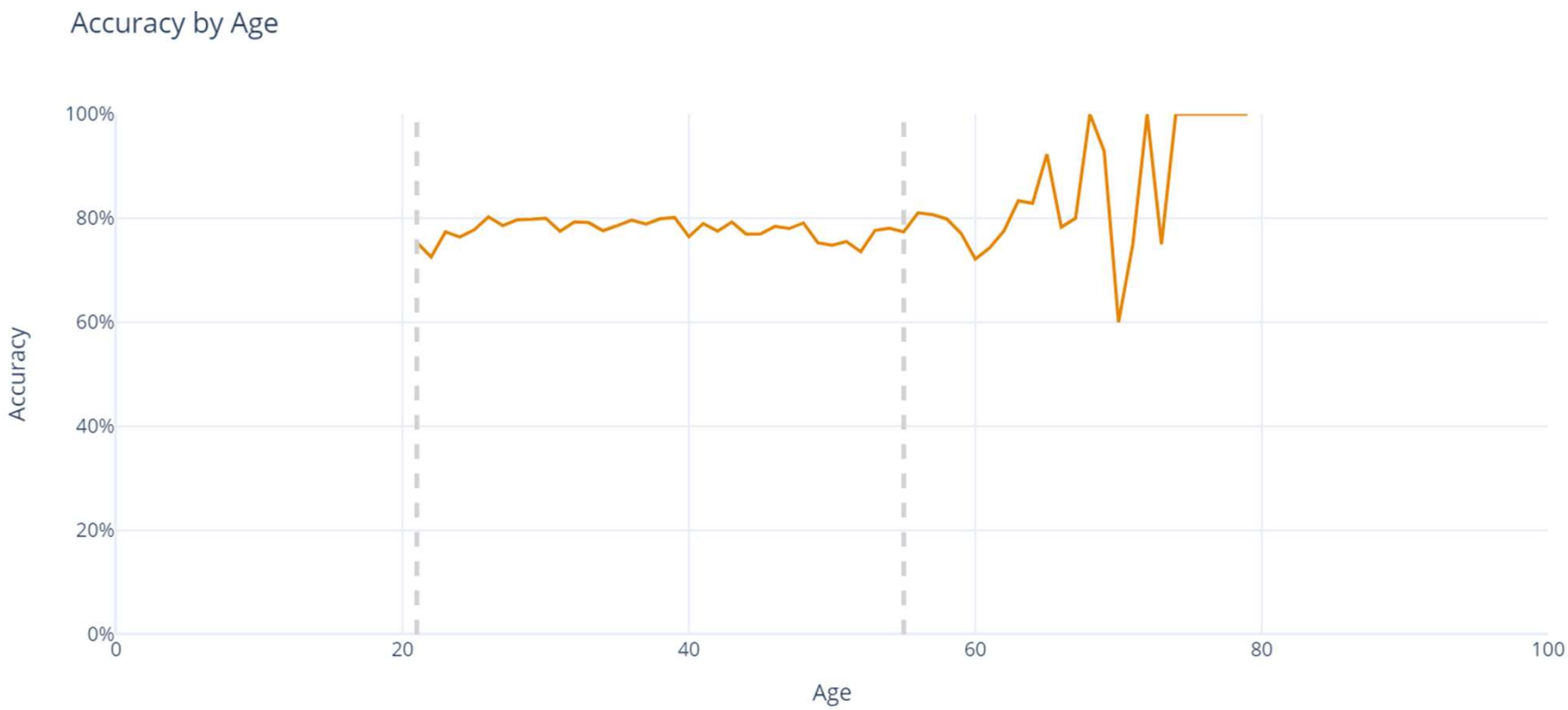




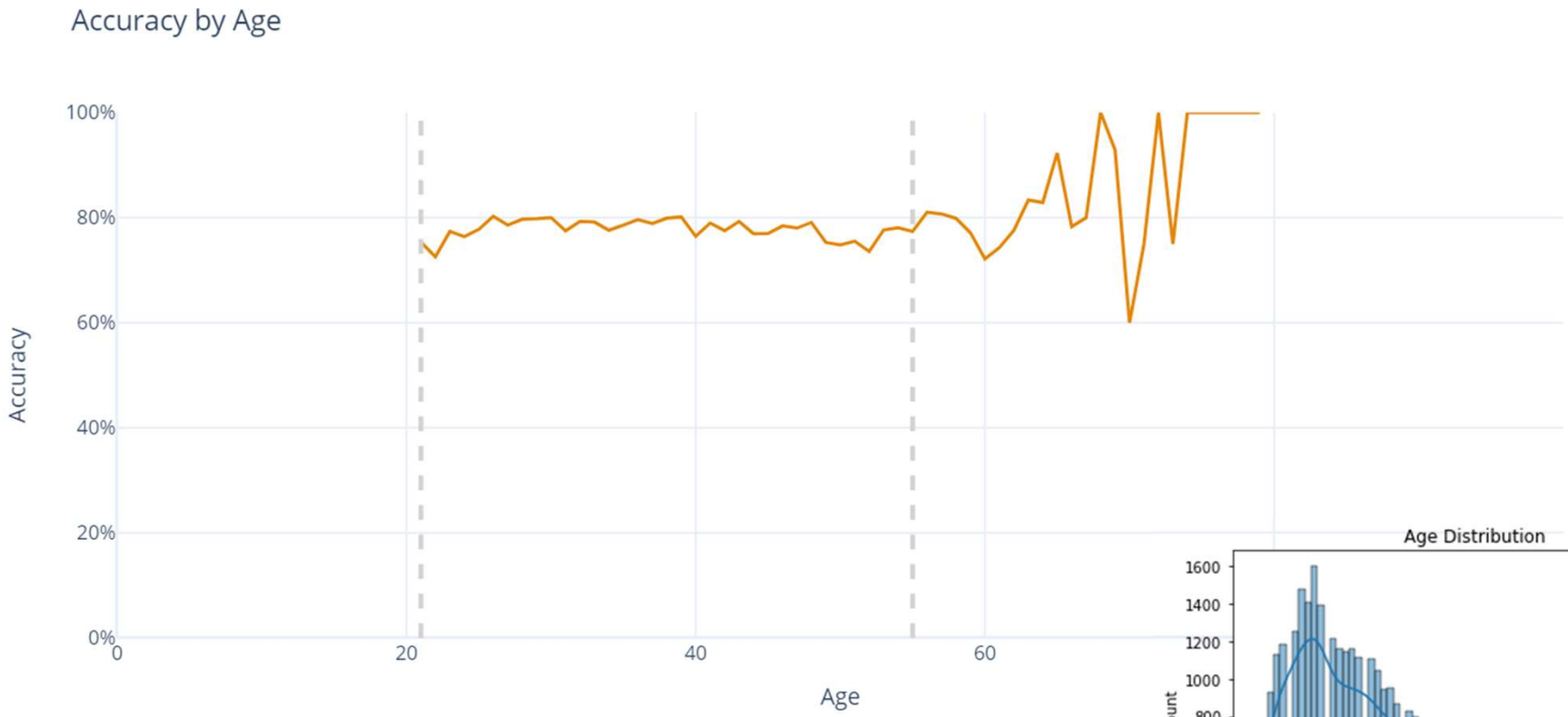
FEATURE IMPORTANCE

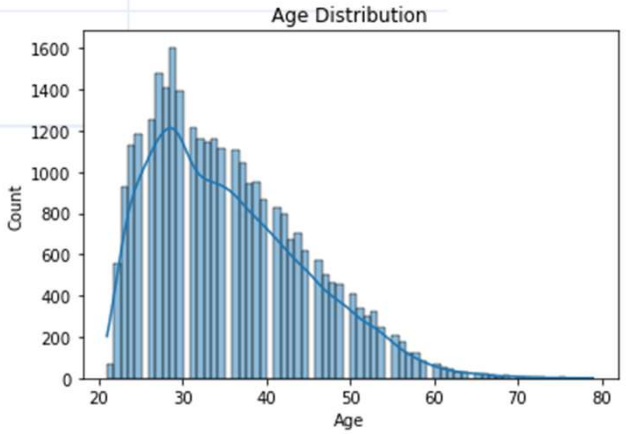
FEATURE IMPORTANCE





MODEL FAIRNESS - AGE

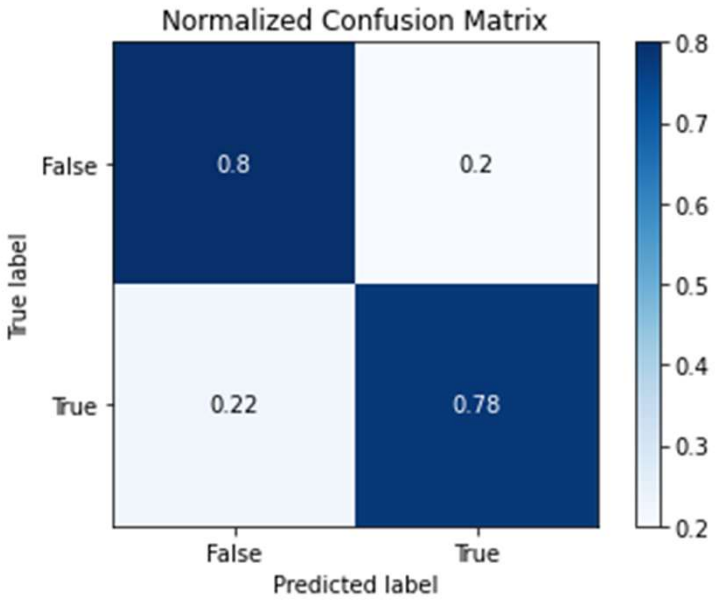




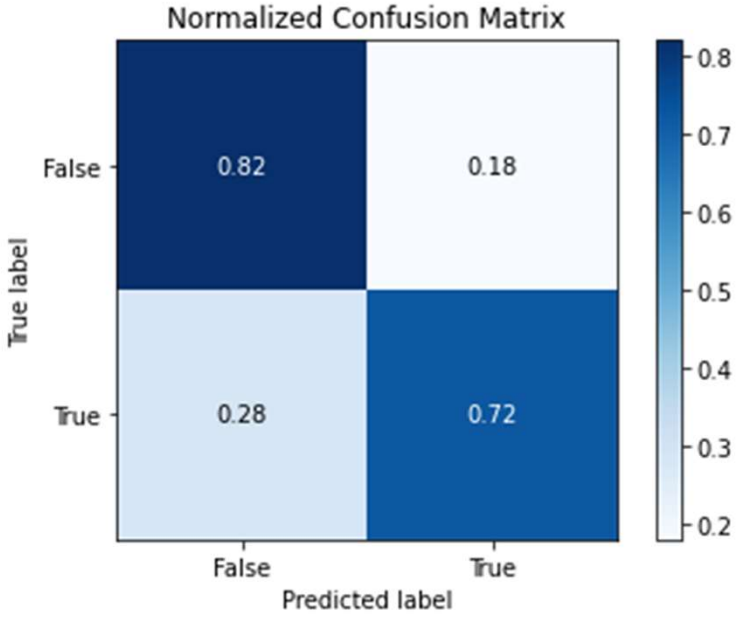
MODEL FAIRNESS - AGE

MODEL FAIRNESS - GENDER

Male



Non-Male



FINAL THOUGHTS