MACHINE LEARNING ALGORITHM APPLCATION FOR PROCESSING OF BANK LOAN APPLICATIONS

By:
Ramkishore Rao
DSA 5900 / Credit Hrs: 4 hrs
Summer 2022
Faculty Advisors: Dr. Trafalis/Dr. Radhakrishnan
Faculty Coordinator: Dr. Beattie

TABLE OF CONTENTS

List of Tables, forthcoming

Preliminary tables included now

List of Exhibits, forthcoming

Preliminary exhibits included now

List of Appendices, forthcoming

Preliminary appendix included now

1.0 Introduction

This project serves as my final practicum for my master's degree in Data Science and Analytics being completed at the University of Oklahoma. As part of this project, various machine learning algorithms were applied to a bank loan dataset[1] to aid in the processing of loan applications from consumers at a bank. For this study, a git hub repository developed by Dr. Jeff Heaton for his Deep Learning (DL)[2] class at Washington University at St. Louis and his accompanying book[3] were leveraged. In addition, class notes from Dr. Nicholson and from Dr. Diochnos were also utilized during the study.

The primary programming language used was Python, with its pre-existing modules. PostgreSQL was used for storing the data. Tableau has been used during the initial exploration phase of the data.

2.0 Objectives

The main objective of the project is to use the existing bank loan dataset to develop back-end statistics models in order to provide a decision on the loan applications. Training, validation, and testing were performed using the existing dataset. An implementation plan is provided below.

3.0 Exploratory Data Analysis

A bank loan dataset[1] that contained 112 features was utilized in this study. Of the 112 features, one of the features was default_date, i.e., this feature had the data on which default occurred. This feature was the target class, and if default had occurred, it was assigned a value of 1 and if default had not occurred, it was assigned a value of 0.

More explanation to follow in final deliverable.

Table 1: Data Breakdown by Target Class

Overall Class Counts
Defaulted: 1
Not Defaulted: 0

Target Class	
0	156,588
1	80,635
Grand Total	237,223

Count of Target Class broken down by Target Class.

[1] : Loan Dataset file from https://www.bondora.com/en/public-reports
[2] https://github.com/jeffheaton/t81_558_deep_learning
[3] Applications of Deep Neural Networks with Keras, Jeff Heaton, Fall 2022.0

3.1 Analysis Summary

A few tables and exhibits are provided in the following pages. They presented breakout of aggregated values of several features by target class value (i.e., 0 if debtor has not defaulted and 1 if debtor has defaulted). Further explanation to be provided in final deliverable.

Table 2: Income Breakouts by Target Class

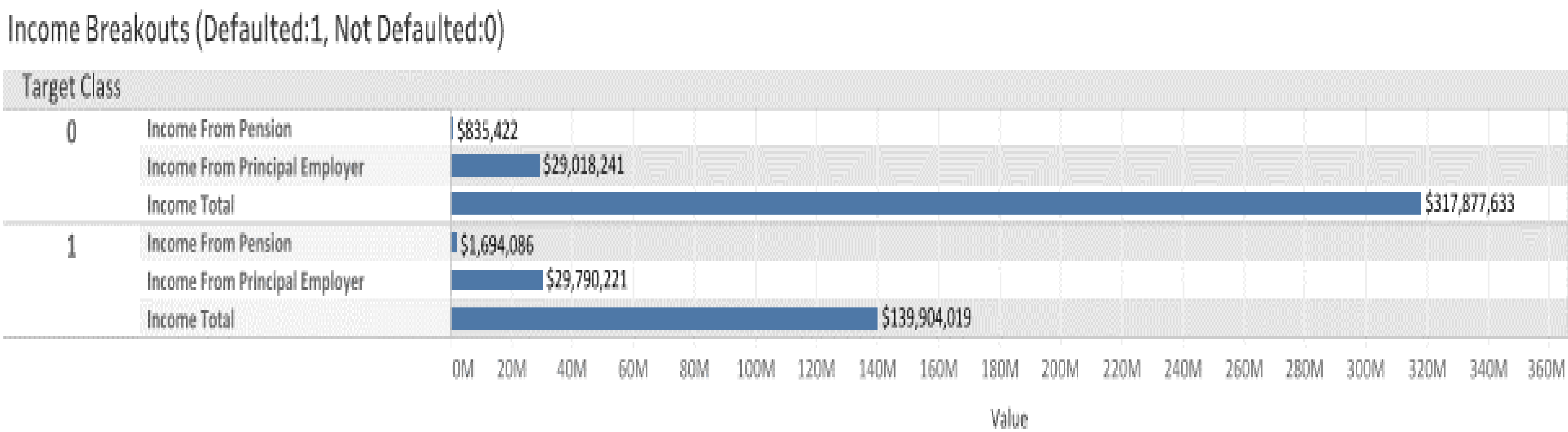


Table 3: Interest Servicing Breakouts by Target Class

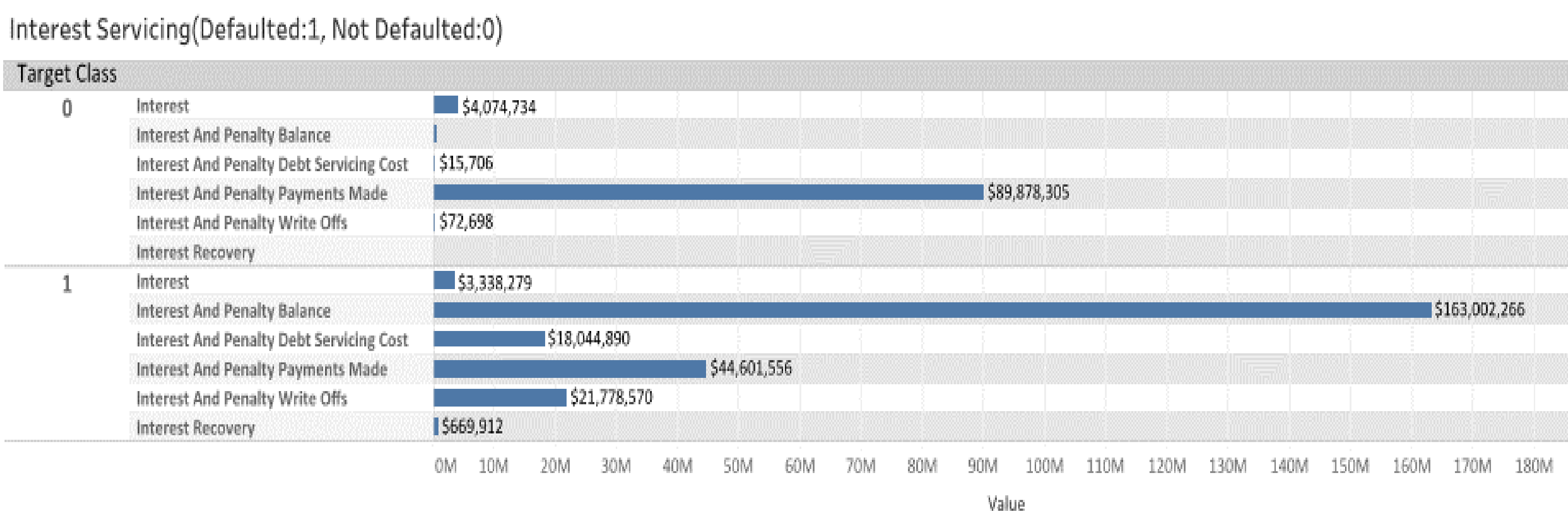


Table 4: Liability Breakouts by Target Class

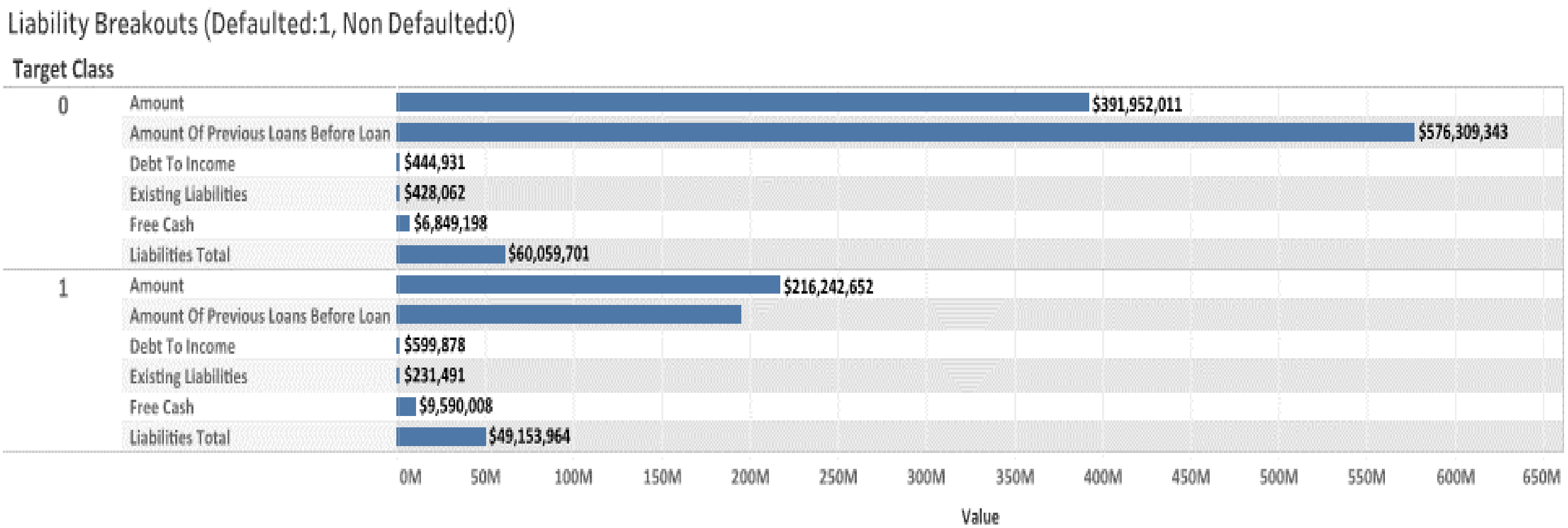


Table 5: Credit Rating by Median Probability of Default

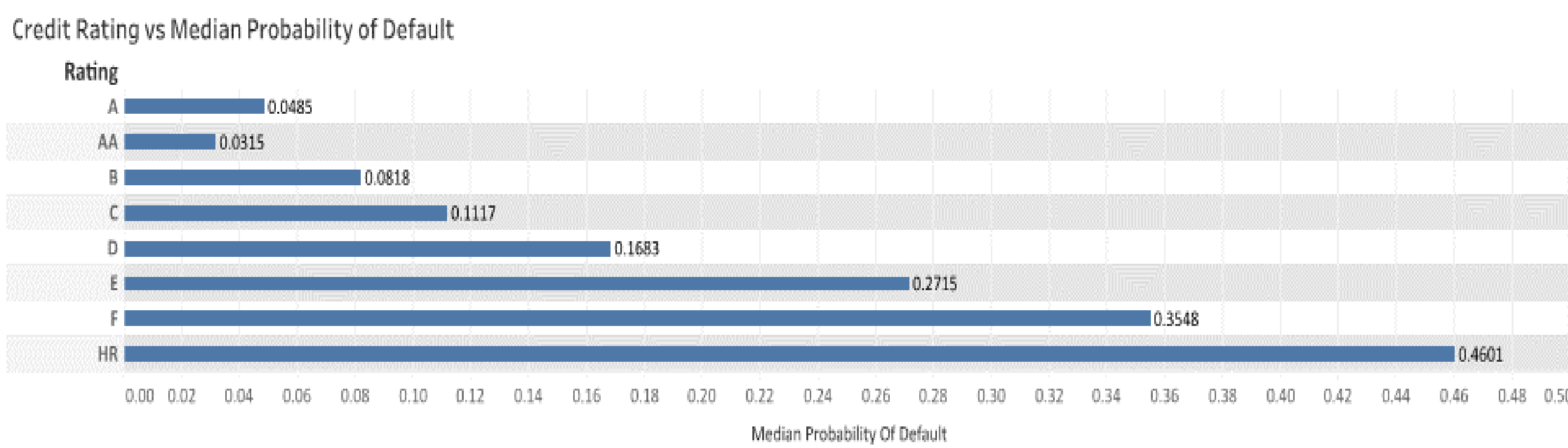


Table 6: Employment Status Counts Breakdown by Target Class

Employment Status
Defaulted: 1
Not Defaulted: 0

Target Class	Employment Status							
	-1	0	2	3	4	5	6	Grand Total
0	140,054	5	456	13,782	428	1,147	595	156,467
1	60,581	27	728	16,278	875	860	1,205	80,554
Grand Total	200,635	32	1,184	30,060	1,303	2,007	1,800	237,021

Count of Employment Status broken down by Employment Status vs. Target Class. The view is filtered on Employment Status, which excludes Null.

Table 7: Work Experience/Home Ownership Type Counts Breakdown by Target Class

Work Experience/Home Ownership Category Breakouts
Defaulted: 1
Not Defaulted: 0

Target Class	Home Ownership Type	Work Experience					
		2-5 Yrs	5-10 Yrs	10-15 Yrs	15-25 Yrs	<2 Yrs	>25 Yrs
0	0			2		1	1
	1	394	941	917	1,369	205	1,567
	2	629	742	421	287	242	103
	3	436	608	407	348	204	248
	4	226	333	256	209	62	193
	5	15	23	36	31	7	58
	6	108	173	62	97	57	54
	7	162	285	244	326	100	232
	8	105	306	418	545	65	337
	9	18	36	63	96	3	76
	Total	2,093	3,447	2,826	3,308	946	2,869
1	0	8	8	3	5	2	8
	1	483	891	1,077	1,578	194	1,589
	2	872	1,106	728	598	341	209
	3	615	752	685	594	249	410
	4	322	533	458	438	103	484
	5	36	64	76	91	19	106
	6	144	183	150	119	53	86
	7	147	221	201	263	63	216
	8	73	207	355	532	29	418
	9	5	32	51	84	7	52
	Total	2,705	3,997	3,784	4,302	1,060	3,578

Table 8: Education/Country Type Counts Breakdown by Target Class

Education/Country Breakout Categories
Defaulted: 1
Not Defaulted: 0

Education	Country	Target Class	
		0	1
-1	EE	201	2
	ES		2
	FI	2,048	185
	Total	2,249	189
0	EE		8
	Total		8
1	EE	12,718	4,819
	ES	460	1,650
	FI	5,869	2,878
	Total	19,047	9,347
2	EE	2,079	2,490
	ES	131	654
	FI	288	798
	SK		4
	Total	2,498	3,946
3	EE	18,943	7,073
	ES	677	2,087
	FI	23,756	10,516
	SK	1	35
	Total	43,377	19,711
4	EE	44,575	17,282
	ES	2,592	7,265
	FI	5,687	3,713
	SK	13	175
	Total	52,867	28,435
5	EE	20,076	5,569

Table 9: Amount of Previous Credit Breakdown by Target Class

Amount of Previous Credit Breakout
Defaulted: 1
Not Defaulted: 0

No Of Previous Loans Before Loan	Target Class		
	0	1	Grand Total
0	66,782	43,481	110,263
1	32,686	16,216	48,902
2	19,268	8,562	27,830
3	11,713	4,557	16,270
4	7,580	2,611	10,191
5	5,117	1,677	6,794
6	3,532	1,100	4,632
7	2,526	772	3,298
8	1,875	523	2,398
9	1,386	378	1,764
10	1,006	244	1,250
Grand Total	153,471	80,121	233,592

Table 10: Days to Payments Percentage of Total Breakdown by Target Class

Days to Payments Percentage of Total by Target Class
Defaulted: 1
Non Defaulted: 0

Active Late Category	Target Class		
	0	1	Grand Total
0-7	95.84%	4.16%	100.00%
8-15	97.51%	2.49%	100.00%
16-30	86.07%	13.93%	100.00%
31-60	82.02%	17.98%	100.00%
61-90	60.72%	39.28%	100.00%
91-120	33.15%	66.85%	100.00%
121-150	4.34%	95.66%	100.00%
151-180	2.94%	97.06%	100.00%
180+	0.85%	99.15%	100.00%

3.2 Analysis Findings

To be Explained More Thoroughly in Final Deliverable.

When aggregated, following can be ascertained from preliminary data analysis:

1) Higher income appears to result in lower default
2) Higher interest appears to result in higher default
3) Poorer credit rating was assigned a higher median probability of default – a decision variable used by the loan processors prior to processing of the loan
4) Higher previous credit obtained appears to result in lower default
5) Higher education appears to result in lower default
6) More prompt appears to result in lower default

4.0 Feature Evaluation/Extraction

This section does further data exploration:

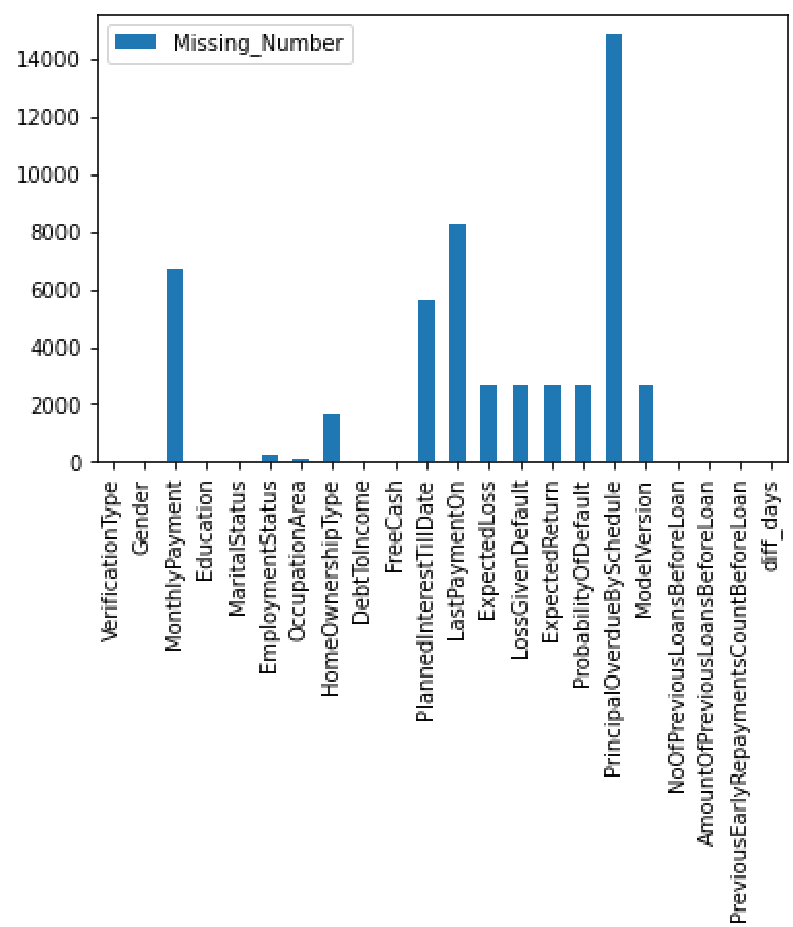
1) Missing value analysis
2) Multi collinearity effects
3) Correlation between predictor variable and target variable
4) PCA analysis to reduce the high dimensional data set – need to evaluate further in final analysis. Will determine if this reduction can still result in reasonable prediction power of the models presented in Section 5.0
5) Final report may also include feature engineering to combine select features for better model interpretability and prediction power.

4.1 Missing Value Analysis

Missing values for key predictor variables is provided in Exhibit 1.

Further explanation to follow in final deliverable to help guide the reader in how features were lowered from the original 112 to what's presented herein.

Exhibit 1: Missing Values Count



4.2 Correlation Analysis

Analysis was conducted to assess for multi-collinearity of predictor variables. The predictor variables that have correlation coefficient greater than 0.75 between each other are presented in Table 11.

Table 11: Correlation Coefficients Between Variables

Variable_1	Variable_2	Correlation Coeff
MaritalStatus	DebtToIncome	0.767
DebtToIncome	MaritalStatus	0.767
NoOfPreviousLoansBeforeLoan	AmountOfPreviousLoansBeforeLoan	0.77
AmountOfPreviousLoansBeforeLoan	NoOfPreviousLoansBeforeLoan	0.77
UseOfLoan	MaritalStatus	0.774
MaritalStatus	UseOfLoan	0.774
MaritalStatus	OccupationArea	0.774
OccupationArea	MaritalStatus	0.774
Interest	ProbabilityOfDefault	0.785
ProbabilityOfDefault	Interest	0.785
EmploymentStatus	DebtToIncome	0.787
DebtToIncome	EmploymentStatus	0.787
AppliedAmount	MonthlyPayment	0.79
MonthlyPayment	AppliedAmount	0.79
UseOfLoan	EmploymentStatus	0.791
EmploymentStatus	UseOfLoan	0.791
EmploymentStatus	OccupationArea	0.791
OccupationArea	EmploymentStatus	0.791
Interest	ExpectedLoss	0.799
ExpectedLoss	Interest	0.799
ExpectedLoss	ProbabilityOfDefault	0.858
ProbabilityOfDefault	ExpectedLoss	0.858
MaritalStatus	EmploymentStatus	0.928
EmploymentStatus	MaritalStatus	0.928
AppliedAmount	Amount	0.947
Amount	AppliedAmount	0.947

In the final analysis, I will evaluate if variables that exhibit multi-collinearity need to be eliminated for models to have better interpretability and prediction power.

Table 12: Correlation Coefficients Between Variables and Target Variable

Variable_Name	Defaulted
Rating_C	-0.182
Status_Repaid	-0.175
Rating_B	-0.136
AmountOfPreviousLoansBeforeLoan	-0.120
PrincipalPaymentsMade	-0.118
NoOfPreviousLoansBeforeLoan	-0.117
ModelVersion	-0.108
LossGivenDefault	-0.098
Rating_D	-0.080
Rating_AA	-0.070
EmploymentDurationCurrentEmployer_UpTo5Years	-0.067
EmploymentDurationCurrentEmployer_Other	-0.049
diff_days	-0.035
Country_FI	-0.032
MonthlyPaymentDay	-0.029
LoanDuration	-0.016
InterestAndPenaltyPaymentsMade	-0.011
LiabilitiesTotal	0.005
EmploymentDurationCurrentEmployer_UpTo1Year	0.005
PreviousEarlyRepaymentsCountBeforeLoan	0.013
EmploymentDurationCurrentEmployer_Retiree	0.013
IncomeFromLeavePay	0.019
Education	0.020
IncomeOther	0.032
HomeOwnershipType	0.033
EmploymentDurationCurrentEmployer_TrialPeriod	0.035
Amount	0.041
Country_SK	0.045
IncomeFromChildSupport	0.046
IncomeFromSocialWelfare	0.046
ExistingLiabilities	0.049
Restructured_True	0.068
AppliedAmount	0.075
EmploymentDurationCurrentEmployer_UpTo4Years	0.076
IncomeFromFamilyAllowance	0.082
FreeCash	0.084
IncomeFromPension	0.085

Table 12 Continued: Correlation Coefficients Between Variables and Target Variable

Variable_Name	Defaulted
EmploymentDurationCurrentEmployer_UpTo3Years	0.091
NewCreditCustomer_True	0.102
EmploymentDurationCurrentEmployer_UpTo2Years	0.108
PrincipalBalance	0.111
RefinanceLiabilities	0.119
Rating_E	0.120
IncomeFromPrincipalEmployer	0.144
MonthlyPayment	0.160
PlannedInterestTillDate	0.187
OccupationArea	0.237
DebtToIncome	0.245
Rating_HR	0.249
UseOfLoan	0.254
Rating_F	0.256
ExpectedReturn	0.273
ActiveScheduleFirstPaymentReached_True	0.277
MaritalStatus	0.282
EmploymentStatus	0.286
Country_ES	0.298
Interest	0.354
ExpectedLoss	0.409
ProbabilityOfDefault	0.432
PrincipalOverdueBySchedule	0.487
Status_Late	0.758
Defaulted	1.000

Final Summary

This shows that the correlation is not significant (less than 0.5, except for Status_Late) between retained variables and the target class variable.

In final deliverable, will remove Status_Late from the predictor variable set, and will explain this analysis in more detail.

This section will attempt to connect the “aggregated” breakout summaries presented in Section 2.0 with the correlation analysis presented herein.

Also, there are some variables may need to be grouped together as part of feature engineering to evaluate if prediction power of the models is better.

4.3 PCA Analysis

Preliminary PCA analysis was conducted to perform exploratory analysis and to evaluate whether the variance in the target class values can be explained by reducing dimensions of the predictor variables.

Initial analysis was conducted using only 5,000 dataset points. This number will be increased prior to the final deliverable.

Initial analysis indicates that 50% of the variance can be explained with 5 principal components (see Exhibit 2).

Separability in the target class is not clearly discernable when 3 principal components are evaluated (see Exhibit 3.

Exhibit 2: Explained Variance vs Principal Component No.

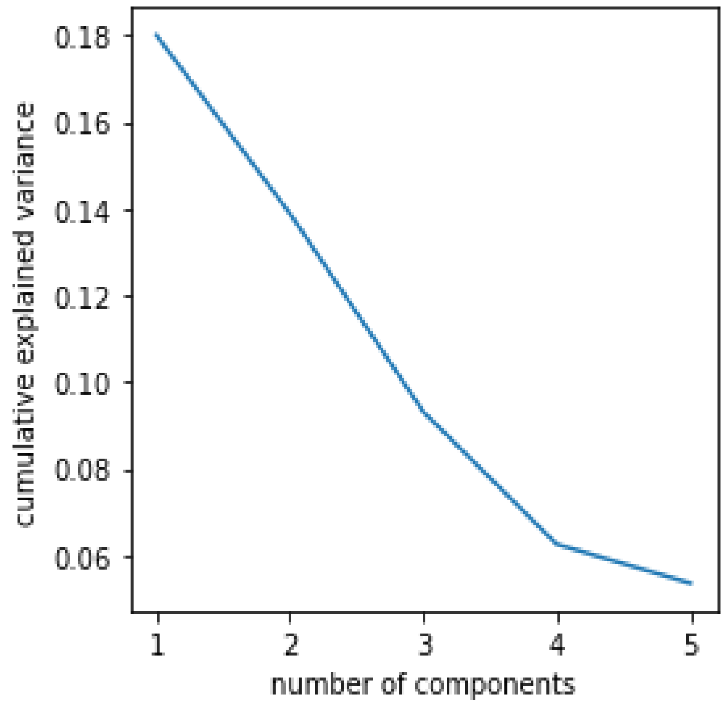
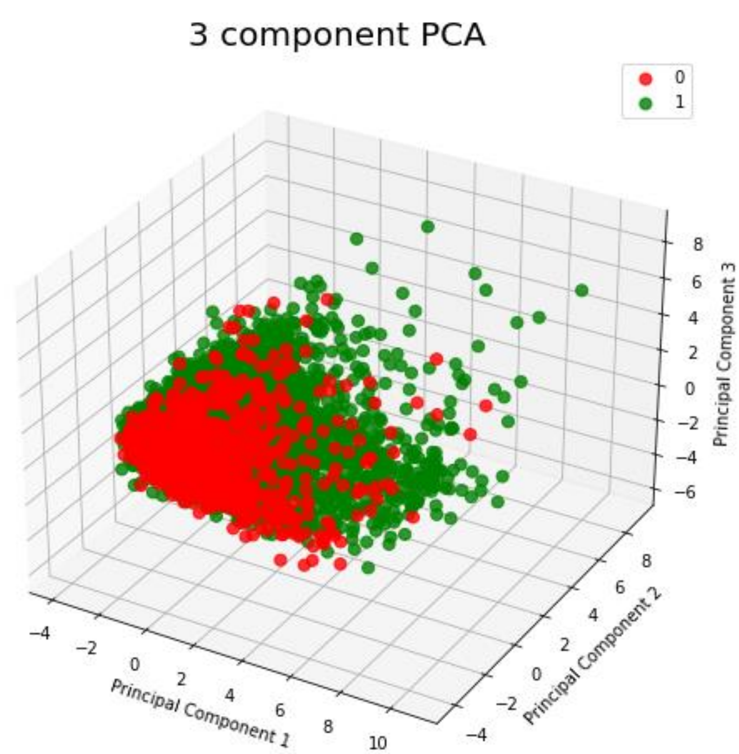


Exhibit 3: Target Class Separation from

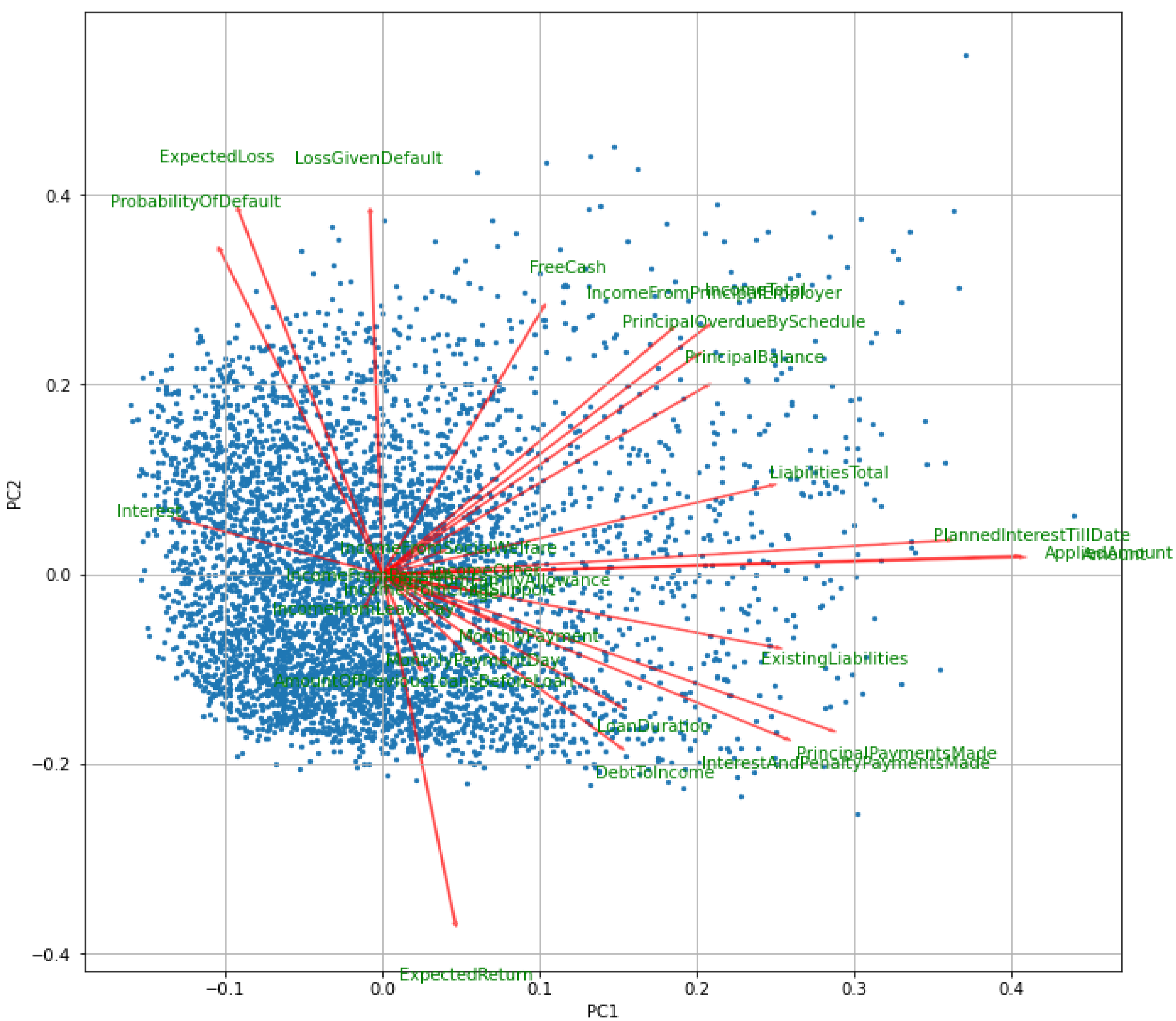
3 Principal Components



PCAA Bi Plot results from this preliminary analysis is presented on Exhibit 4. Based on the "vector" representation of some of the features, it does appear that the first two components may be a reasonable assimilator of some of the predictor variables.

This will be further evaluated during the final analysis to be included in the final deliverable.

Exhibit 4: PCA Bi Plot



5.0 Machine Learning Modeling

Pythion's sklearn and tensorflow/keras is utilized for this analysis. Explanation to follow in final deliverable. Tensorflow federated with keras will also be checked in the final analysis.

Data preparation included the following steps:

a. Prior to further evaluation, the predictor variables were segregated into continuous or categorical variables.
b. The continuous predictor variables were scaled to ensure that the range of input values for the various variables are similar.
c. One hot encoding was utilized to encode the categorical variables as either dummy or ordinal[4].

Dataset was split into train, validation, and test components using sklearn. The weights for the train, validation, and test components are being developed. At this time, results from the test dataset following training on the train set is presented.

Final analysis will include testing on the validation data set and will also include results from 5-fold validation prior to testing on the test dataset.

Model optimization will be performed to tune hyperparameters in each type of model for the final analysis. Typical hyperparameters for the various models are as follows.

i. Logistic Regression with its variants, include lasso, ridge regression, and elastic net – parameters for training may include L1 and L2 regularization parameters, solver, or class_weights [5].
ii. Naïve Bayes - parameter will be alpha, Additive (Laplace/Lidstone) smoothing parameter (0 for no smoothing and 1 for default)[6].
iii. K nearest neighbors – parameter will be number of nearest neighbors to be used for model classification.
iv. Decision Tree Classifiers – parameters will include criterion (gini or logloss), splitter (best or random), max_depth (depth of tree), etc[7].
v. ANN, tensorflow and tensorflow federated – parameters may include number of hidden layers, type of activation functions, regularization – L1, L2, dropout, and learning_rate. Tensorflow's in-built Bayesian Optimization will be utilized for hyperparameter tuning[8].

---

[4] Refer to Section 2.2.2 Encoding Categorical Variables as dummies, Applications of Deep Neural Networks with Keras, Jeff Heaton, Fall 2022.0

[5] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

[6] https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html

[7] https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

[8] 8.4 – Bayesian Hyperparameter Optimization for Keras, Applications of Deep Neural Networks with Keras, Jeff Heaton, Fall 2022.0

5.1 Logistic Regression

#Basic Logistic Regression Model Fitting,

w/Default, Explanation to be provided in Final Deliverable

```
log_reg = LogisticRegression()
log_reg.fit(X_train1, y_train1)
y_predict = log_reg.predict(X_train1)
```

Optimization, Hyerparameter Tuning is forthcoming

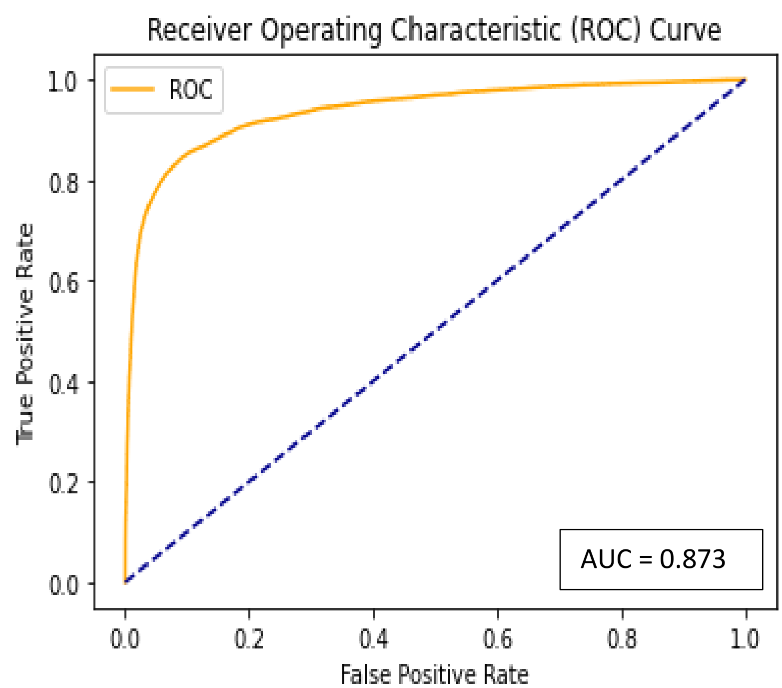
Exhibit 5: Performance Evaluation: Logistic Regression

Confusion Matrix, Test Dataset:

	Predicted No Default	**Predicted Yes Default**
Actual No Default	25,341	1,846
Actual Yes Default	2,721	11,894

Parameter	**Value**
RMSE	0.33
Precision	0.866
Accuracy	0.891
Recall	0.814
F1_Score	0.839

Exhibit 6: ROC Curve: Logistic Regression



5.2 Multinomial Bayes

#Basic MNB Model Fitting,

w/Default, Explanation to be provide in Final Deliverable

```
clf = MultinomialNB()
clf.fit(X_train1, y_train1)
y_predict = clf.predict(X_train1)
```

Optimization, Hyerparameter Tuning is forthcoming

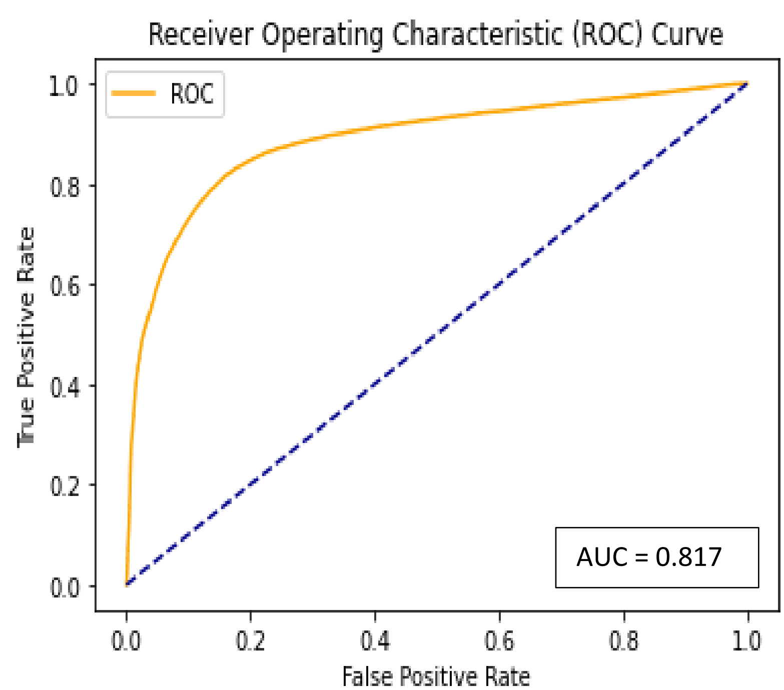
Exhibit 7: Performance Evaluation: Multinomial Bayes

Confusion Matrix, Test Dataset:

	Predicted No Default	**Predicted Yes Default**
Actual No Default	24,278	2,909
Actual Yes Default	3,775	10,840

Parameter	**Value**
RMSE	0.40
Precision	0.788
Accuracy	0.840
Recall	0.741
F1_Score	0.764

Exhibit 8: ROC Curve: Multinomial Bayes



5.3 Decision Tree

5 stumps model, explanation to be provided in final deliverable

```
tree_clf1 =
DecisionTreeClassifier(criterion='entropy',
max_depth = 5)
tree_clf1.fit(X_train1, y_train1)
y_predict = tree_clf1.predict(X_train1)
```

Optimization, Hyerparameter Tuning is forthcoming

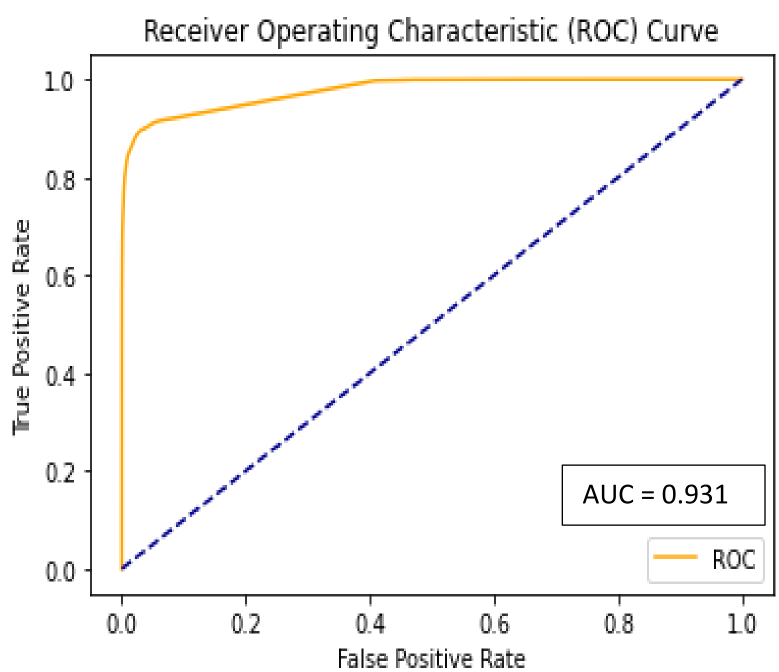
Exhibit 9: Performance Evaluation: Decision Tree

Confusion Matrix, Test Dataset:

	Predicted No Default	**Predicted Yes Default**
Actual No Default	26,544	643
Actual Yes Default	1,684	12,931

Parameter	**Value**
RMSE	0.236
Precision	0.953
Accuracy	0.944
Recall	0.885
F1_Score	0.917

Exhibit 10: ROC Curve: Decision Tree



5.4 Ensemble Trees

Explanation to be provided in final deliverable

```
ada_clf1 = AdaBoostClassifier(DecisionTreeClassifier(max_depth =1), n_estimators = 20)
ada_clf1.fit(X_train1, y_train1)
y_predict = ada_clf1.predict(X_train1)
```

Optimization, Hyerparameter Tuning is forthcoming

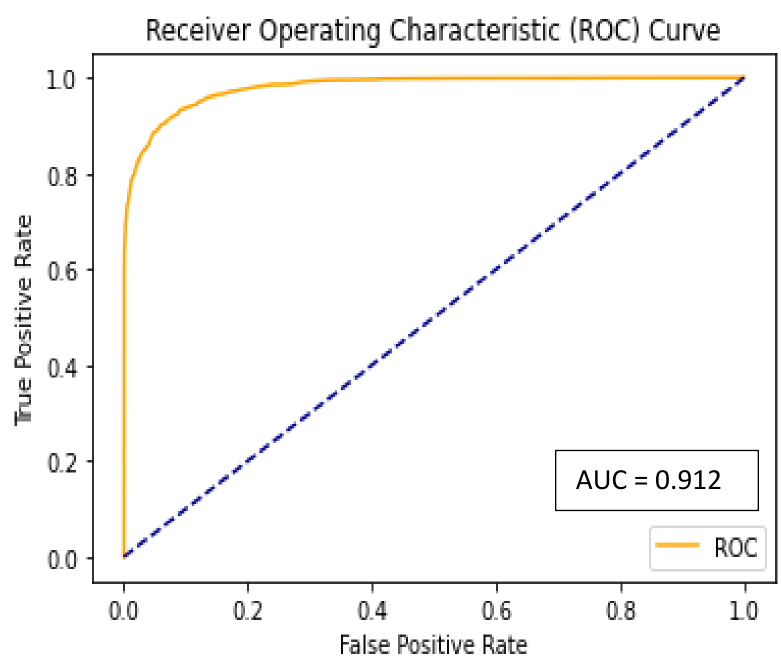
Exhibit 11: Performance Evaluation: Ensembles Trees

Confusion Matrix, Test Dataset:

	Predicted No Default	**Predicted Yes Default**
Actual No Default	26,063	1,124
Actual Yes Default	1,972	12,643

Parameter	**Value**
RMSE	0.272
Precision	0.918
Accuracy	0.926
Recall	0.865
F1_Score	0.891

Exhibit 12: ROC Curve: Ensemble Trees



5.5 Deep Neural Network with Tensorflow/Keras

Explanation to be provided in Final Deliverable

1 input layer, 2 hidden layers, and 1 output layer, RELU activation for all except output – sigmoid activation

```
model = Sequential()
model.add(Dense(100, input_dim=X_train1.shape[1], activation='relu',
            kernel_initializer='random_normal'))
model.add(Dense(50,activation='relu',kernel_initializer='random_normal'))
model.add(Dense(25,activation='relu',kernel_initializer='random_normal'))
model.add(Dense(1,activation='sigmoid',kernel_initializer='random_normal'))
model.compile(loss='binary_crossentropy',
          optimizer=tensorflow.keras.optimizers.Adam(),
          metrics =['accuracy'])
monitor = EarlyStopping(monitor='val_loss', min_delta=1e-3,
   patience=5, verbose=1, mode='auto', restore_best_weights=True)
model.fit(X_train1,y_train1,validation_data=(X_test,y_test),
       callbacks=[monitor],verbose=2,epochs=1000)
```

Optimization, Hyerparameter Tuning is forthcoming

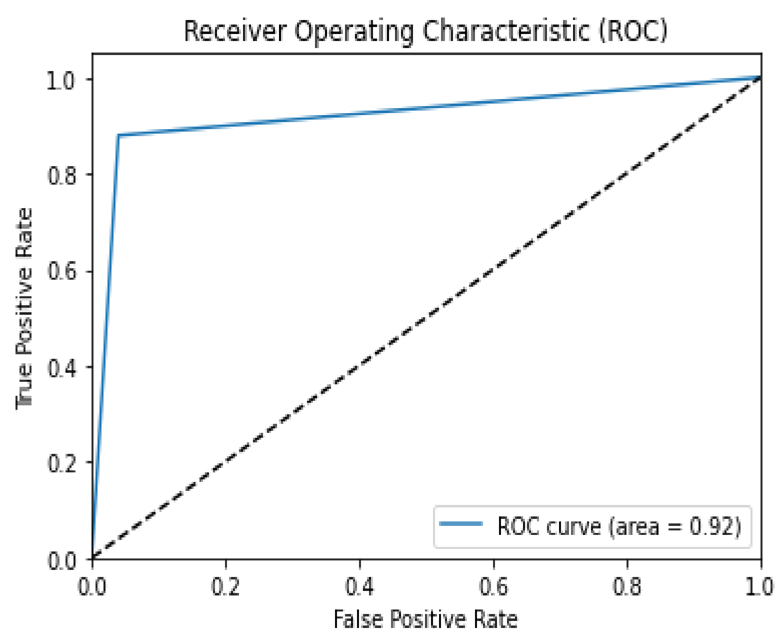
Exhibit 13: Performance Evaluation: Tensor Flow/Keras

Confusion Matrix, Test Dataset:

	Predicted No Default	**Predicted Yes Default**
Actual No Default	26,101	1,086
Actual Yes Default	1,768	12,847

Parameter	**Value**
RMSE	0.261
Precision	0.922
Accuracy	0.931
Recall	0.879
F1_Score	0.900

Exhibit 14: ROC Curve: Tensor Flow/Keras



5.6 Tensorflow Federated/Keras

forthcoming

5.7 Summary of Findings

Forthcoming

Preliminary results are promising, with decision tree and deep neural networks providing the best results. Final analysis will include optimization and hyperparameter tuning.

6.0 Conclusions

forthcoming

7.0 References

Footnote referencing will be replaced with traditional referencing to be included in this section.

APPENDIX

Feature No	Feature Name
1	ReportAsOfEOD
2	LoanId
3	LoanNumber
4	ListedOnUTC
5	BiddingStartedOn
6	BidsPortfolioManager
7	BidsApi
8	BidsManual
9	PartyId
10	NewCreditCustomer
11	LoanApplicationStartedDate
12	LoanDate
13	ContractEndDate
14	FirstPaymentDate
15	MaturityDate_Original
16	MaturityDate_Last
17	ApplicationSignedHour
18	ApplicationSignedWeekday
19	VerificationType
20	LanguageCode
21	Age
22	DateOfBirth
23	Gender
24	Country
25	AppliedAmount
26	Amount
27	Interest
28	LoanDuration
29	MonthlyPayment
30	County
31	City
32	UseOfLoan
33	Education
34	MaritalStatus
35	NrOfDependants
36	EmploymentStatus
37	EmploymentDurationCurrentEmployer
38	EmploymentPosition
39	WorkExperience
40	OccupationArea
41	HomeOwnershipType
42	IncomeFromPrincipalEmployer

43	IncomeFromPension
44	IncomeFromFamilyAllowance
45	IncomeFromSocialWelfare
46	IncomeFromLeavePay
47	IncomeFromChildSupport
48	IncomeOther
49	IncomeTotal
50	ExistingLiabilities
51	LiabilitiesTotal
52	RefinanceLiabilities
53	DebtToIncome
54	FreeCash
55	MonthlyPaymentDay
56	ActiveScheduleFirstPaymentReached
57	PlannedPrincipalTillDate
58	PlannedInterestTillDate
59	LastPaymentOn
60	CurrentDebtDaysPrimary
61	DebtOccuredOn
62	CurrentDebtDaysSecondary
63	DebtOccuredOnForSecondary
64	ExpectedLoss
65	LossGivenDefault
66	ExpectedReturn
67	ProbabilityOfDefault
68	PrincipalOverdueBySchedule
69	PlannedPrincipalPostDefault
70	PlannedInterestPostDefault
71	EAD1
72	EAD2
73	PrincipalRecovery
74	InterestRecovery
75	RecoveryStage
76	StageActiveSince
77	ModelVersion
78	Rating
79	EL_V0
80	Rating_V0
81	EL_V1
82	Rating_V1
83	Rating_V2
84	Status
85	Restructured
86	ActiveLateCategory

87	WorseLateCategory
88	CreditScoreEsMicroL
89	CreditScoreEsEquifaxRisk
90	CreditScoreFiAsiakasTietoRiskGrade
91	CreditScoreEeMini
92	PrincipalPaymentsMade
93	InterestAndPenaltyPaymentsMade
94	PrincipalWriteOffs
95	InterestAndPenaltyWriteOffs
96	PrincipalBalance
97	InterestAndPenaltyBalance
98	NoOfPreviousLoansBeforeLoan
99	AmountOfPreviousLoansBeforeLoan
100	PreviousRepaymentsBeforeLoan
101	PreviousEarlyRepaymentsBefoleLoan
102	PreviousEarlyRepaymentsCountBeforeLoan
103	GracePeriodStart
104	GracePeriodEnd
105	NextPaymentDate
106	NextPaymentNr
107	NrOfScheduledPayments
108	ReScheduledOn
109	PrincipalDebtServicingCost
110	InterestAndPenaltyDebtServicingCost
111	ActiveLateLastPaymentCategory
112	Target Class: Defaulted