

DRAFT

Machine Learning Consumer Loan Processing

Ram Rao

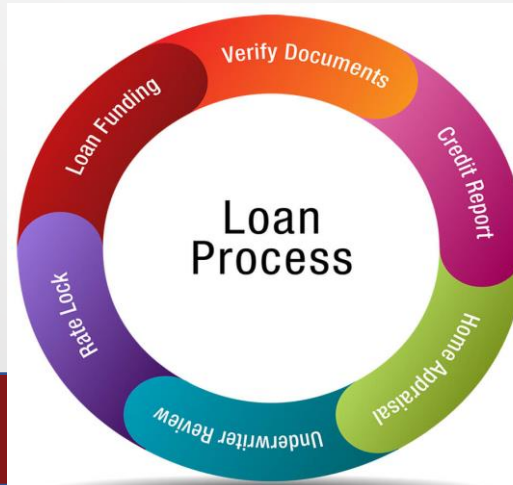
July 1, 2022

DSA 5900 Practicum



Project Definition

- Identify Credit-Worthiness of Loan Applicants at Financial Institutions
 - Apply Machine Learning Models to Evaluate whether Applicants will default on a Loan
- Identify a Process for Remote Machine Learning
- Stakeholders:
 - Agencies that Process Consumer Loans
- Dr. Radhakrishnan and Dr. Trafalis are my advisors



LOAN DOCUMENTATION CHECKLIST



EMPLOYMENT/INCOME

- ☐ Pay stubs for the most recent 30 days available
- ☐ W-2's for the previous two years
- ☐ Federal tax returns for the previous two years. All pages and schedules must be included
- ☐ If self-employed, provide all pages and schedules of last two years' business tax returns and corporate K-1's
- ☐ Proof of additional income, such as Social Security benefits, child support, or alimony (if applicable)

ASSETS

- ☐ Provide ALL pages of most recent 2 months' statements for all accounts; including all checking, savings, stocks, IRA, 401k, etc. The statements must show your name, account number and the name of the banking institution. Any non-payroll deposits will have to be explained and documented.
- ☐ If funds to close will come from a gift, complete the gift letter (will be provided to you) and the following:
 - ☐ From the donor - bank statements showing the funds in the donor's account and a copy of the check from the donor's account
 - ☐ From you - a copy of the deposit slip showing the gift check deposited into your account
- ☐ If funds to close are from sale of home
 - ☐ Estimated closing statement showing anticipated proceeds
 - ☐ Copy of final closing statement and deposit slip showing proceeds deposited into bank account

CREDIT / IDENTIFICATION/ ELIGIBILITY

- ☐ Copy of driver's license or other photo I.D.
- ☐ Copy of divorce decree
- ☐ Copy of bankruptcy papers, including all schedules and discharge, and credit explanation letter for reason for bankruptcy
- ☐ Letter of explanation on any late payments, collections, charge off's or derogatory credit
- ☐ Letter of explanation for all recent credit inquiries
- ☐ If VA, DD214 if not active duty or Statement of service if active duty

PROPERTY

- ☐ Select your insurance agent and provide agent's name, address, and phone number
- ☐ If refinancing, or if you will be retaining your current home or own other property
 - ☐ Current mortgage statement
 - ☐ Copy of insurance declaration page
- ☐ If you're currently renting, provide your Landlord's name, phone number and address. 12 months canceled rent checks will be necessary for private landlords.
- ☐ If you live with a family member, letter stating you live rent-free

POINTmortgage[®] corporation
"A Mortgage Bank"

Data Ingestion



Data Source:

<https://www.bondora.com/en/public-reports>

Tableau, Python, Sckit Learn,
Tensorflow/Keras,
PyTorch and PySft

No of Features

111 Predictor Variables

1 Target Variable

- Defaulted : 1
- Non Defaulted : 0

Overall Class Counts

Defaulted: 1

Not Defaulted: 0

Target Class	Count of Target Class	% of Total Count of Target Class)
0	156,588	66.0%
1	80,635	34.0%
Grand Total	237,223	100.0%

Count of Target Class and % of Total Count of Target Class) broken down by Target Class.

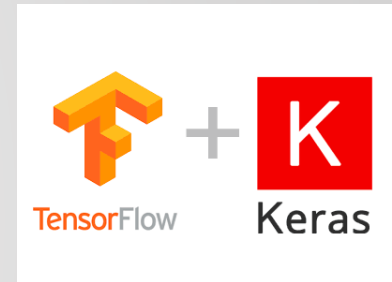


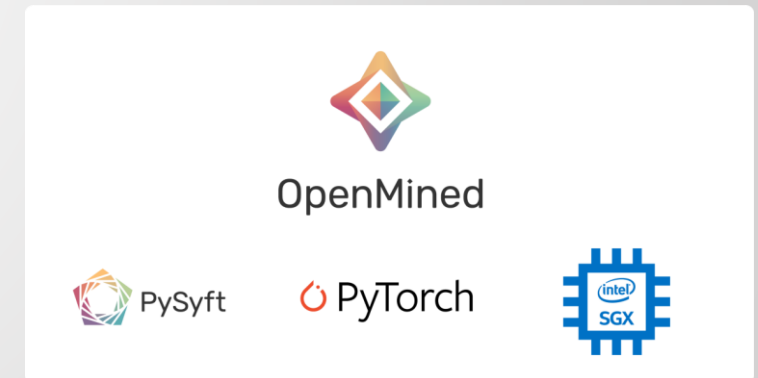
Tableau : Data Viz

Python: Data Processing

Sckit Learn: ML Models

Tensorflow/Keras: Neural Net

PyTorch, PySft: Remote ML



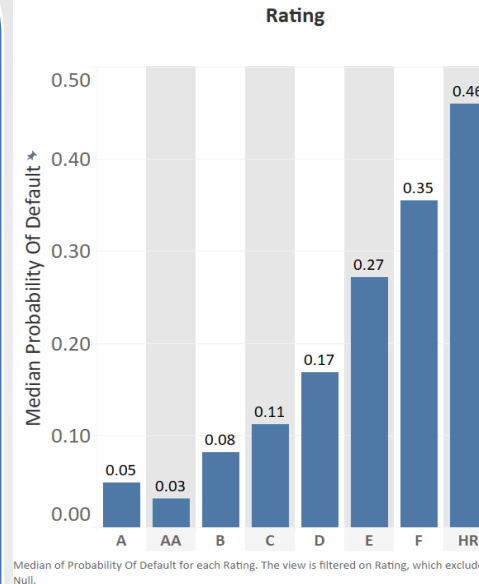
Data Exploration and Preparation - 1



Exploratory Analysis:

- ☐ Lower Default
 - Higher Income
 - Lower Interest Servicing
 - Better Credit Rating
 - Higher Previous Credit
 - Higher Education
 - More Prompt Payment
- ☐ No Significant Multicollinearity
- ☐ Correlation Not High Between Predictor and Target

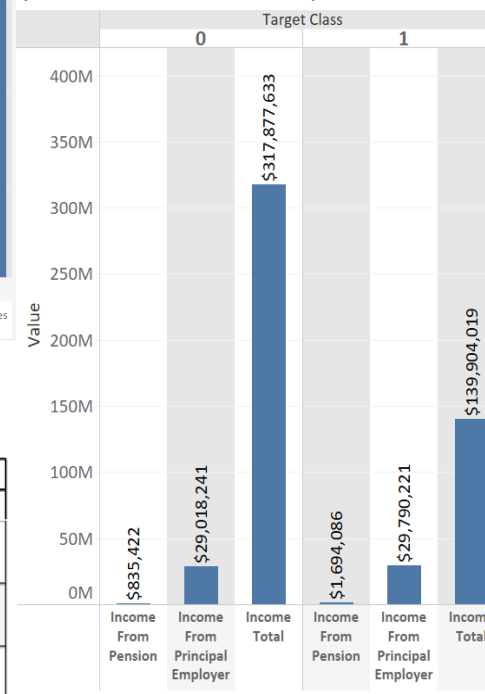
Credit Rating vs Median Probability of Default



Days to Payments By Class Percentage of Total vs Days Outstanding
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%
151-180	2.94%	97.06%	100.00%
180+	0.85%	99.15%	100.00%

Income Breakouts
(Defaulted:1, Not Defaulted:0)



Correlation Coefficient

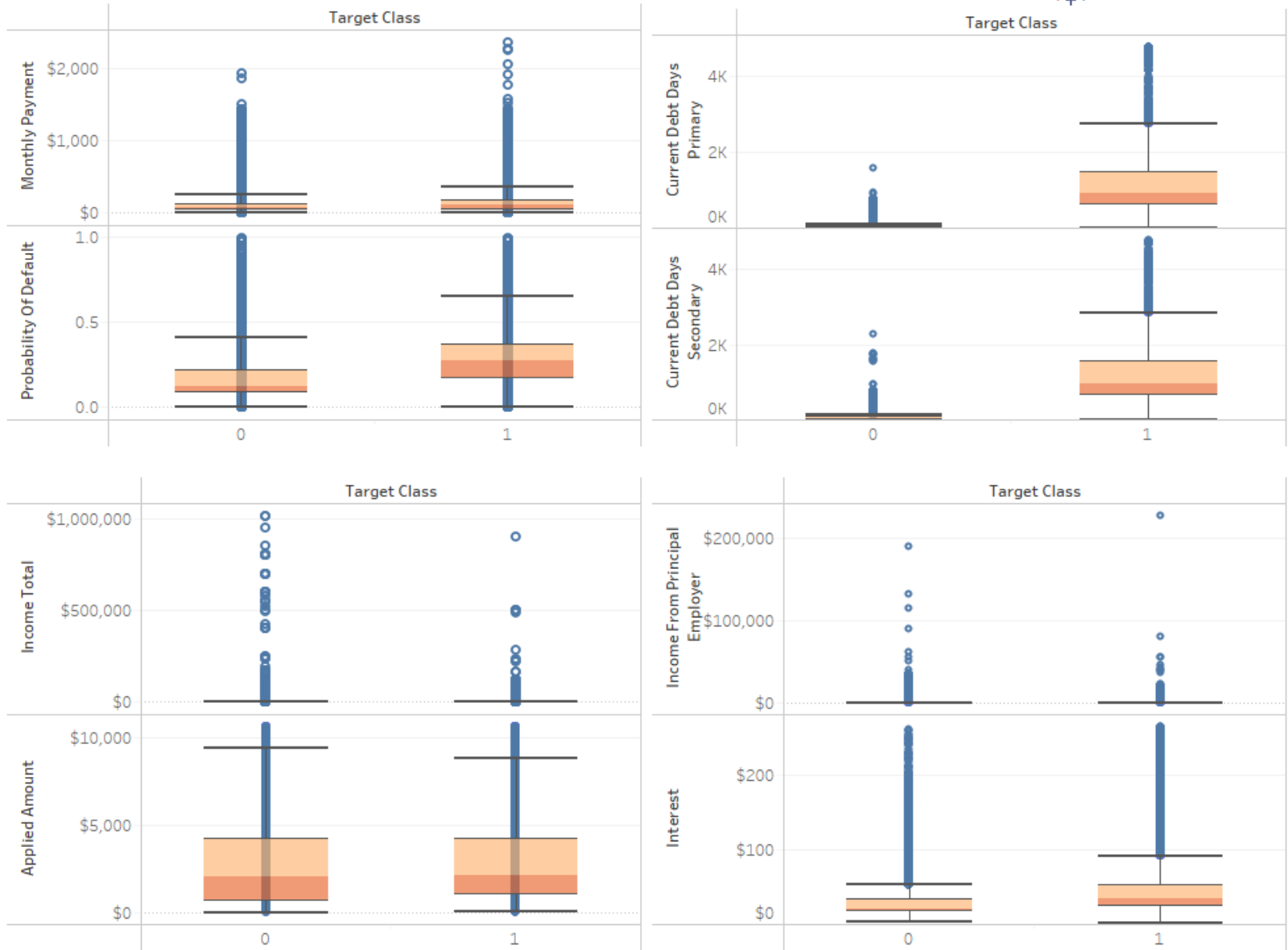
Variable_Name	Defaulted
EmploymentDurationCurrentEmployer_UpTo3Years	0.091
NewCreditCustomer_True	0.102
EmploymentDurationCurrentEmployer_UpTo2Years	0.108
PrincipalBalance	0.111
RefinanceLiabilities	0.119
Rating_E	0.120
IncomeFromPrincipalEmployer	0.144
MonthlyPayment	0.160
PlannedInterestTillDate	0.187
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

Data Exploration and Preparation - 2

Exploratory Analysis:

- ❑ Higher Spread and Max for Target Class 1
 - Probability of Default
 - Debt Types
 - Interest Servicing
- ❑ No Significant Differences Between Classes
 - Applied Amount
 - Income Types
- ❑ Missing Values Eliminated Preliminarily

Box and Whiskers - Predictor Variables

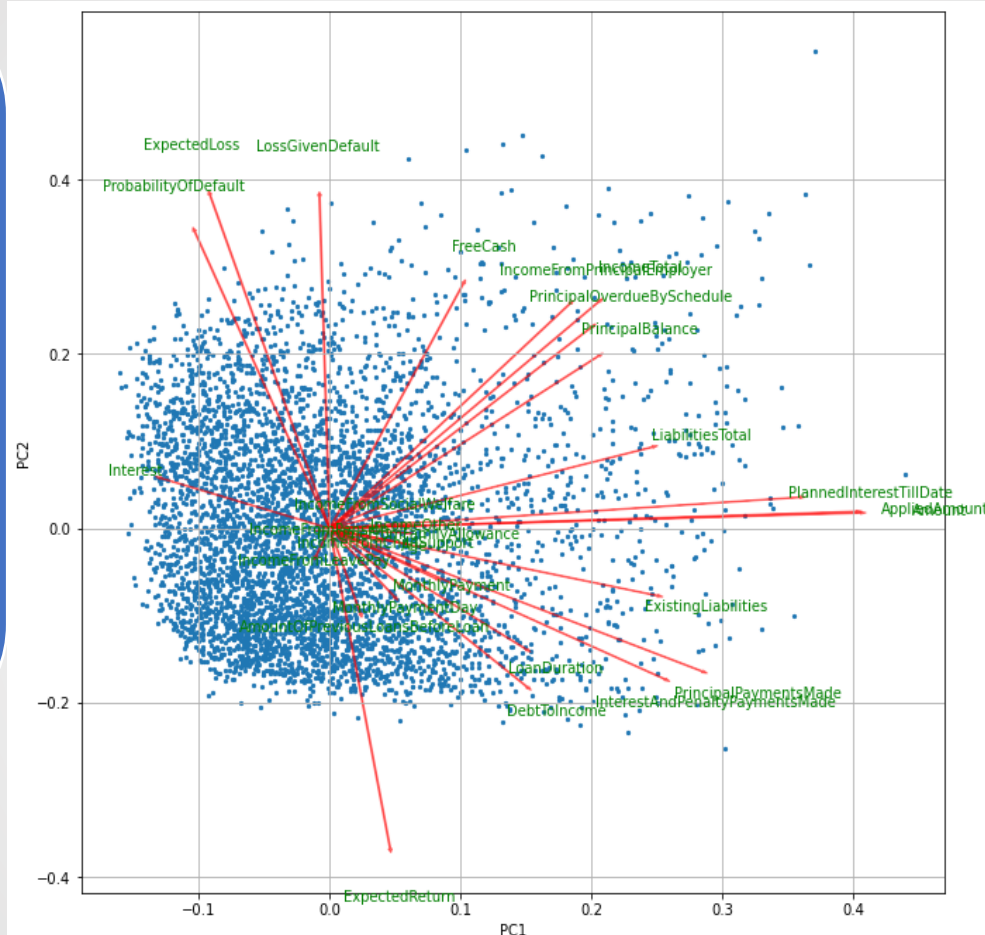
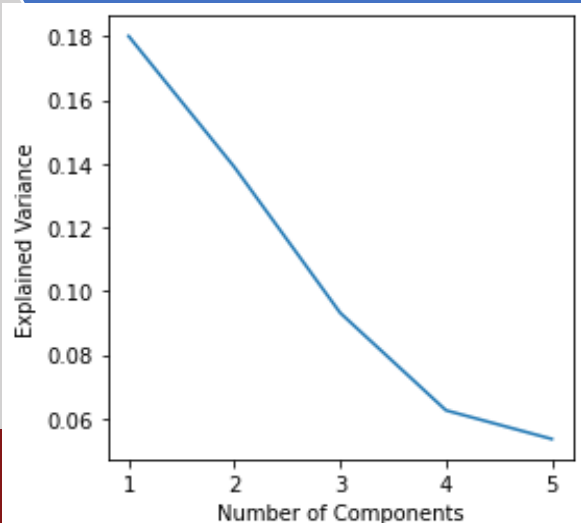


PCA Assessment



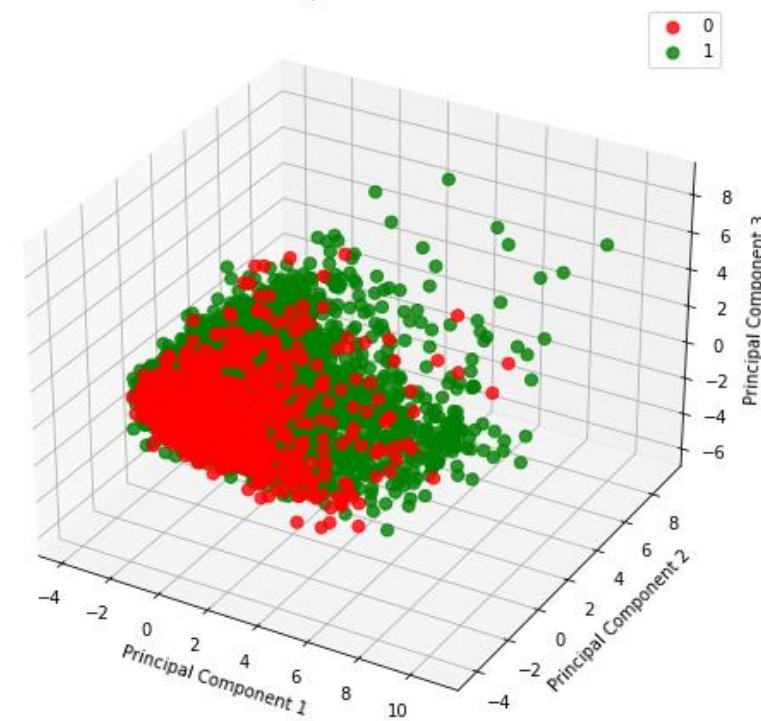
PCA Analysis:

- ✓ 5,000 Dataset Points Analyzed
- ✓ No of Continuous Variables Scaled and Transformed: 28
- ✓ Limited Variance Explained by 5 Components
- ✓ No Significant Separation Between Classes Observed from PCA 1, 2, and 3
- ✓ Bi Plot shows Explanation of Few Features from PCA 1 and 2

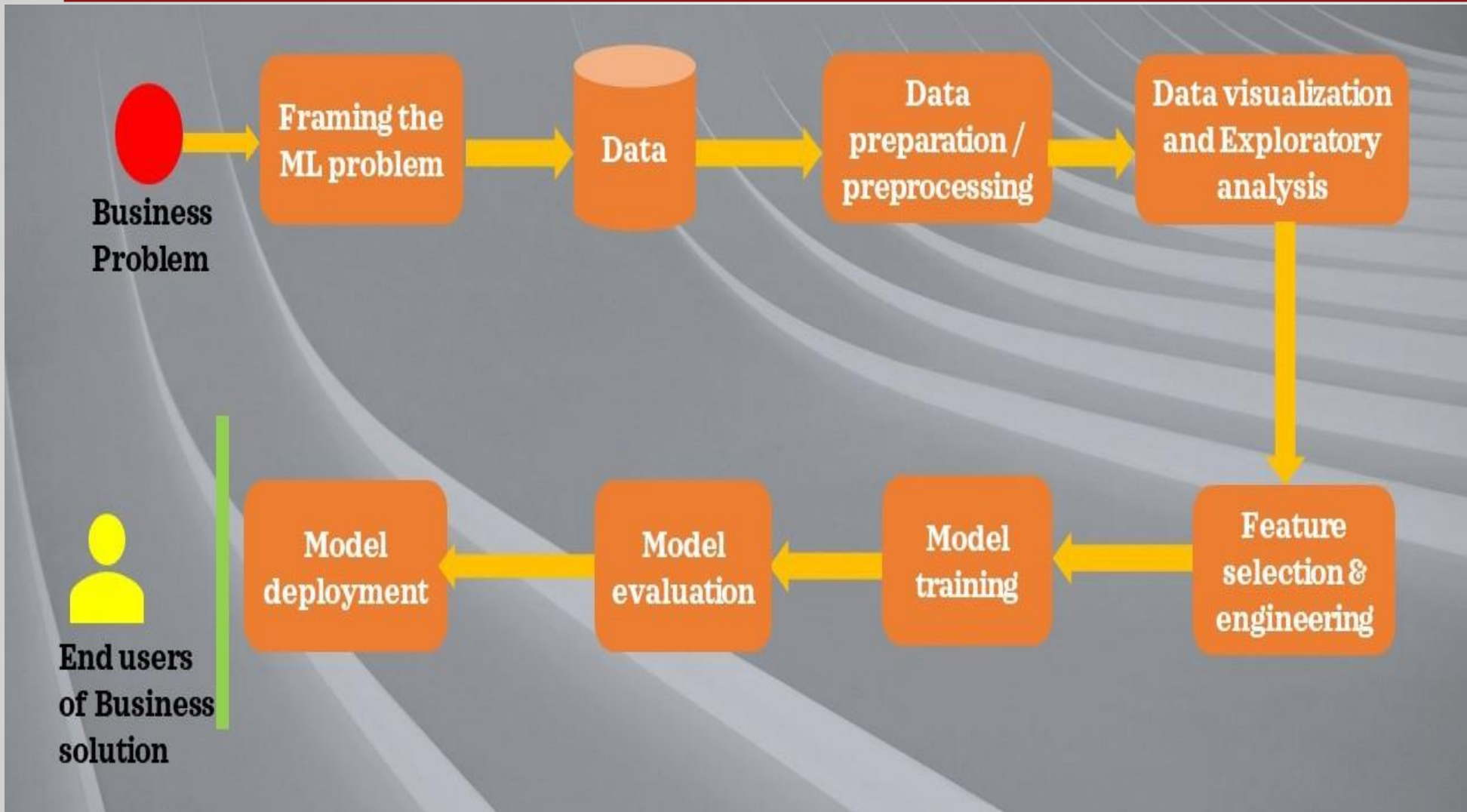


PCA Not a Significant Benefit to Model Predictability, Categorical Count Outweighs Continuous Variables

3 component PCA



Modeling Preprocessing And Overview



☐ Preprocessing with Sckit-Learn

- ✓ Scaled Continuous Variables
- ✓ One hot encoded Categorical Variables

☐ Modeling, Training/Testing

- ✓ Sckit Learn
- ✓ Tensorflow Keras
 - Default
 - GridSearch CV Optimization
- ✓ Remote Machine Learning – PyTorch and PySft
- ✓ Sckit-Learn Metrics for Evaluation

Model Results - Logistic Regression and Naïve Bayes

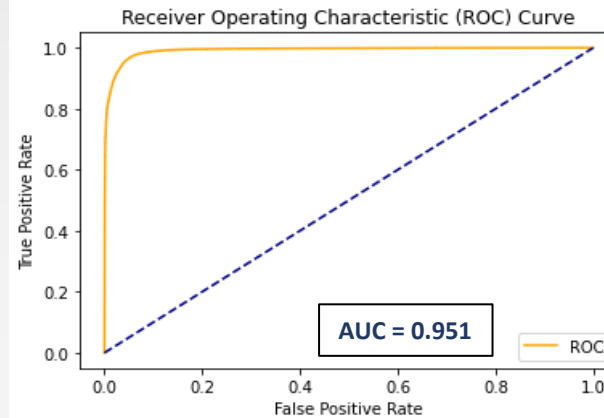


Logistic Regression:

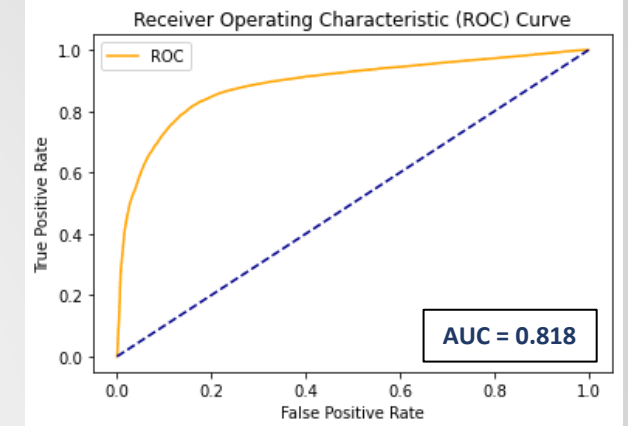
- Grid Search 5-Fold CV
- 200 Iterations
- Hyperparameters
 - ✓ Penalty: L1 and L2
 - ✓ C : 1, 5, 10
 - ✓ Solver, lbfgs, liblinear and saga

Naïve Bayes:

- Grid Search 5-Fold CV
- Hyperparameters
 - ✓ Alpha: 1E-4, 1E-2, 1E-1, and 1



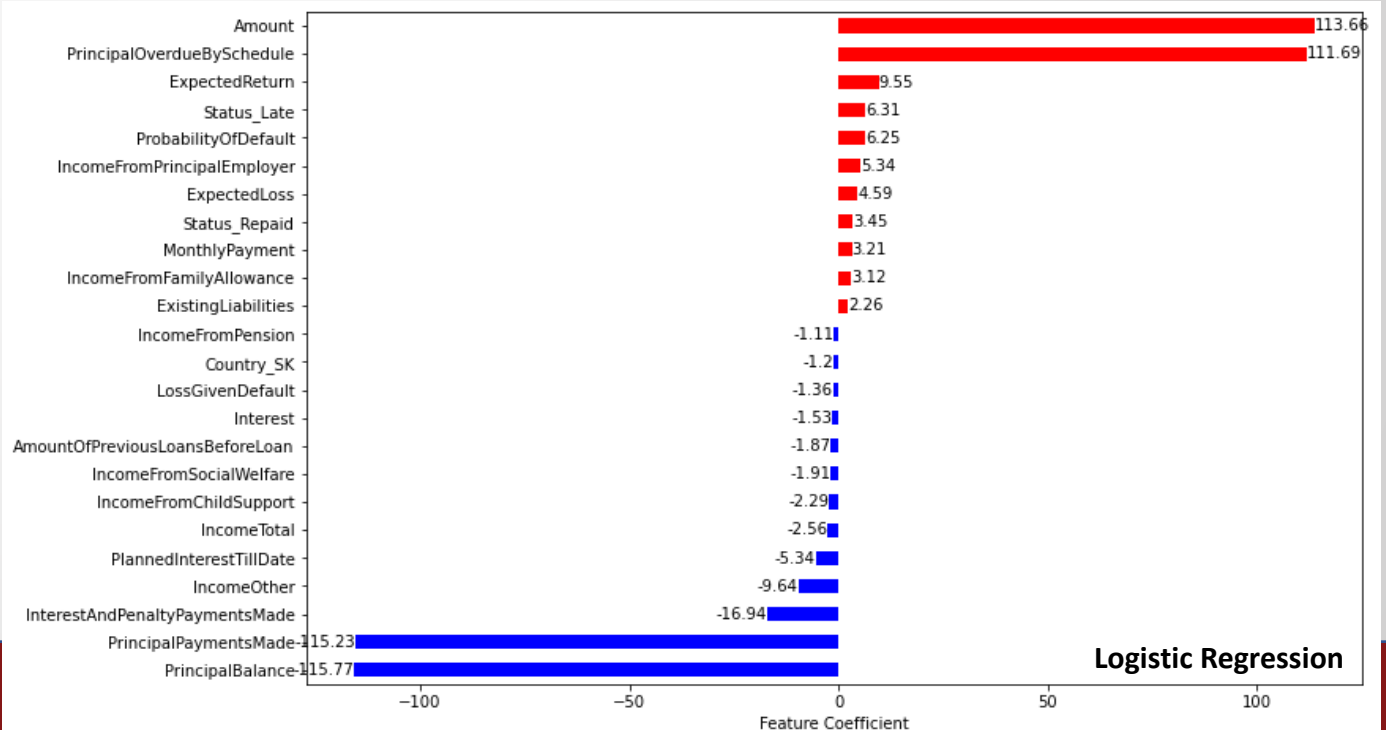
Logistic Regression



Naïve Bayes

Logistic Regression	Class 0 Predicted	Class 1 Predicted
Class 0 Actual	26,280	907
Class 1 Actual	928	13,687

Naïve Bayes	Class 0 Predicted	Class 1 Predicted
Class 0 Actual	24,283	2,904
Class 1 Actual	3,762	10,853



Logistic Regression

Model Results - Decision Trees and Ensemble Forests

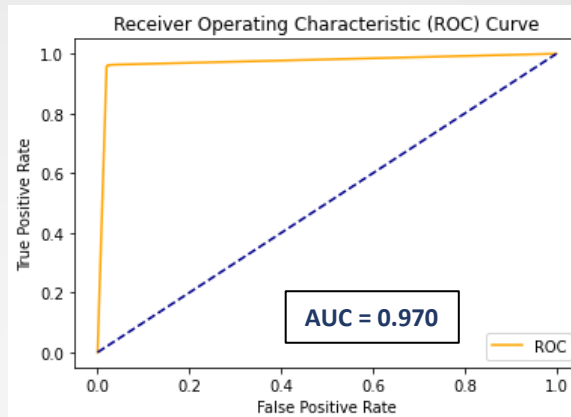


Decision Trees:

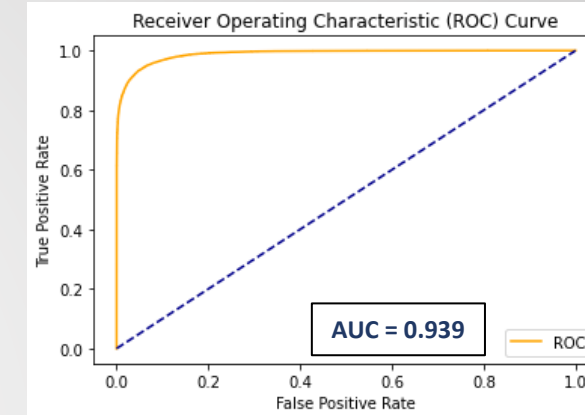
- Grid Search 5-Fold CV
- Hyperparameters
 - ✓ Criterion : gini, entropy
 - ✓ Max_depth : 5,10,20

Ensemble Forests:

- Grid Search 5-Fold CV
- Hyperparameters
 - ✓ N_estimators: 5,10,20, 50, 100
 - ✓ Learning_Rate: 0.1, 0.5, 1.0, 2.0, 5.0



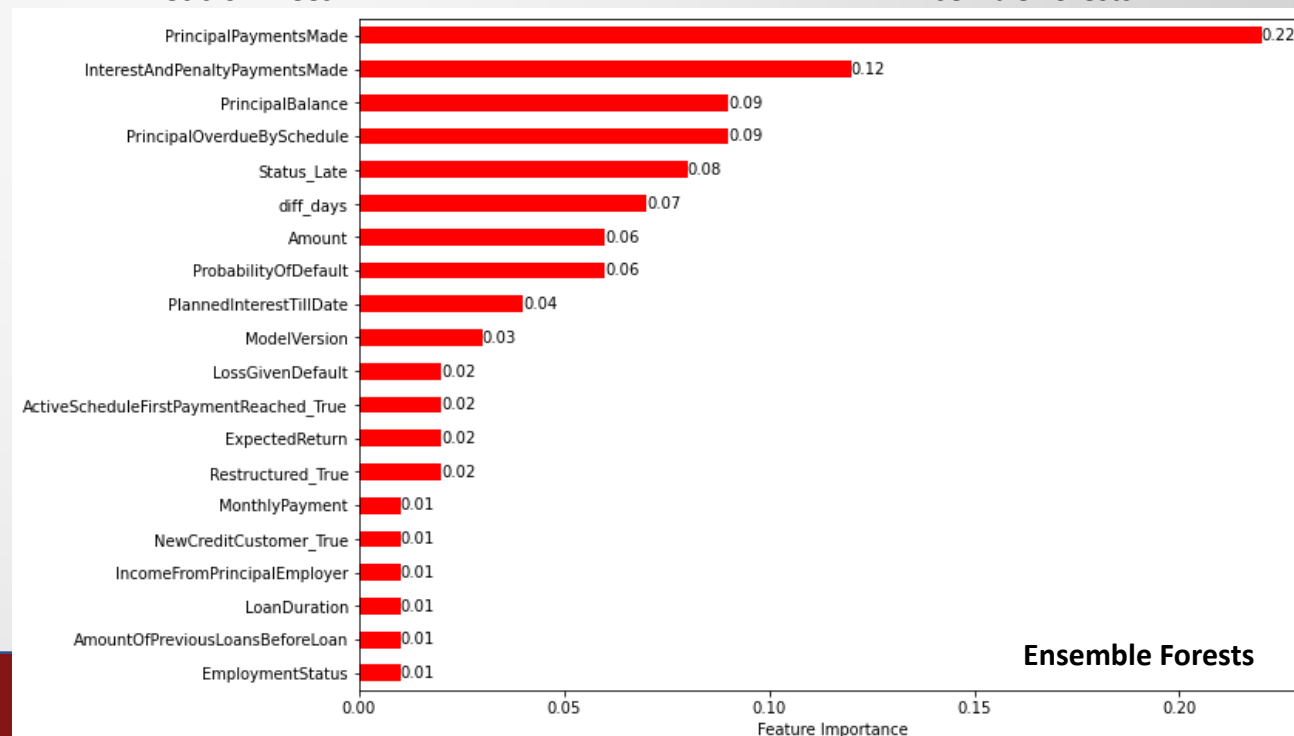
Decision Trees



Ensemble Forests

Decision Trees	Class 0 Predicted	Class 1 Predicted
Class 0 Actual	26,663	554
Class 1 Actual	591	14,024

Ensemble Forests	Class 0 Predicted	Class 1 Predicted
Class 0 Actual	26,238	949
Class 1 Actual	1,276	13,339



Ensemble Forests

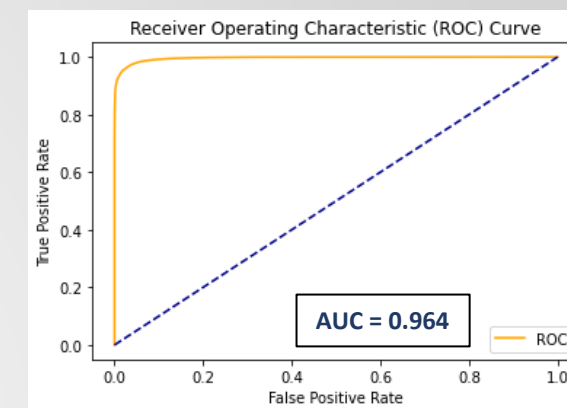
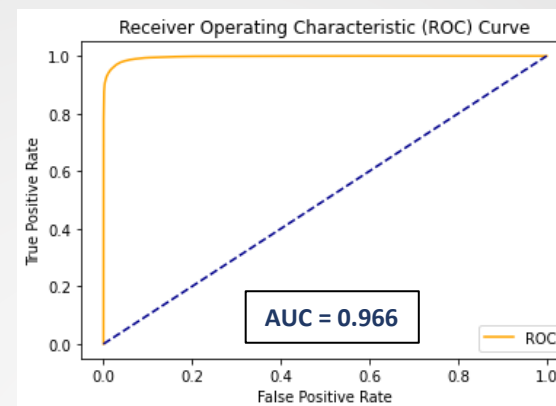
Model Results Random Forests



Random Forests:

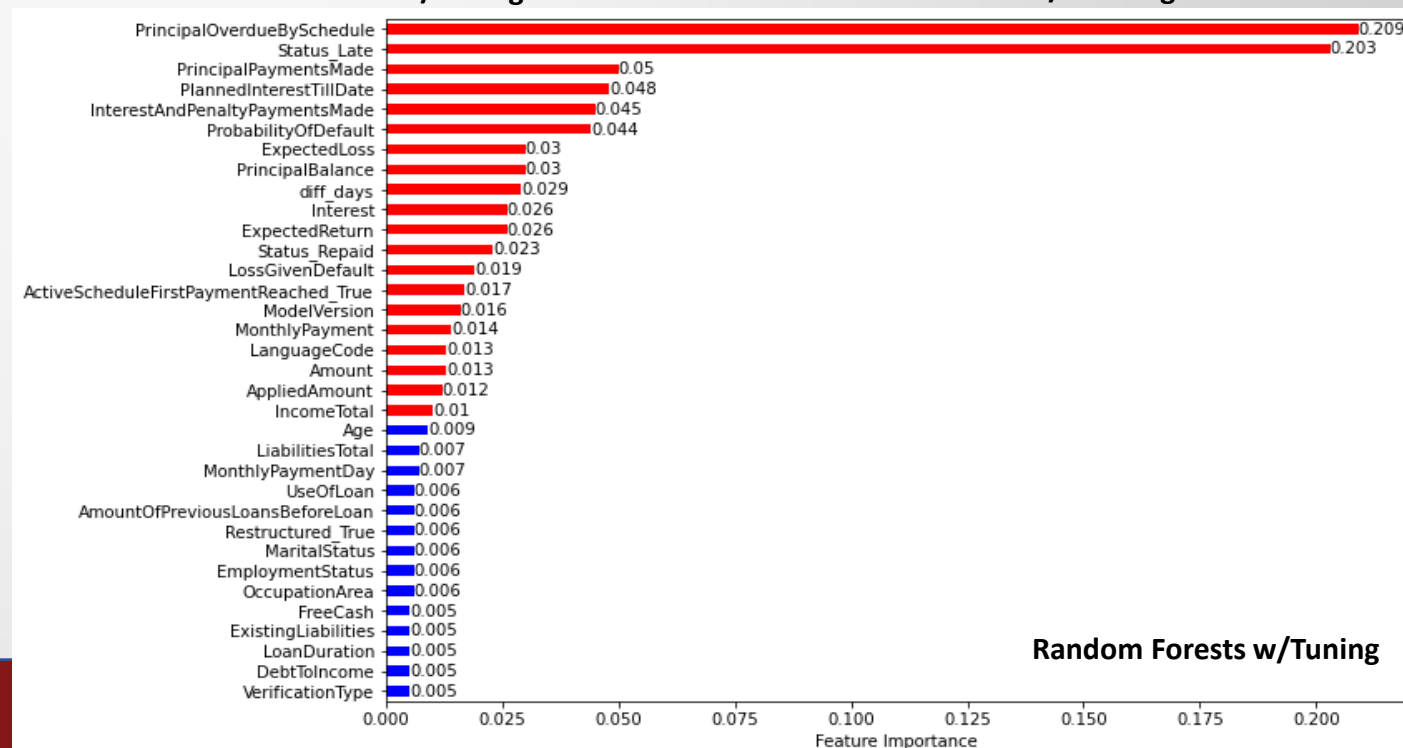
- Grid Search 5-Fold CV
- Hyperparameters
 - ✓ N_estimators: 50, 100, 200
 - ✓ Criterion: gini, entropy
 - ✓ Max_features: sqrt, log2, auto

Random Forests	Class 0 Predicted	Class 1 Predicted
Class 0 Actual	26,854	333
Class 1 Actual	826	13,789



Random Forests w/Tuning

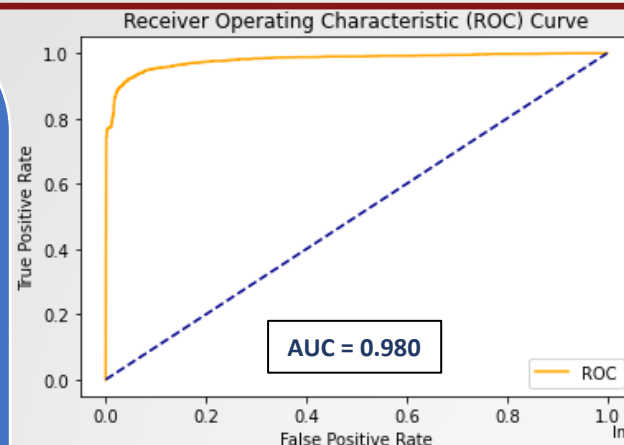
Random Forests w/o Tuning



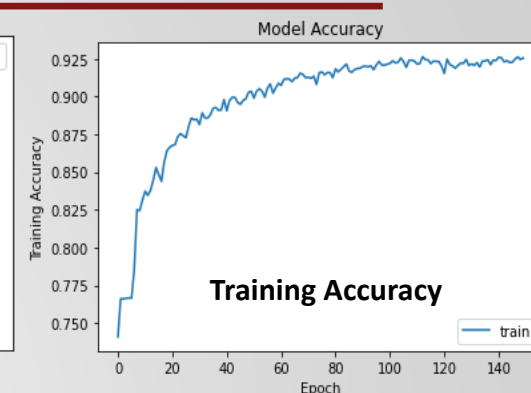
Model Results - Neural Nets, Keras/Tensorflow

Neural Net:

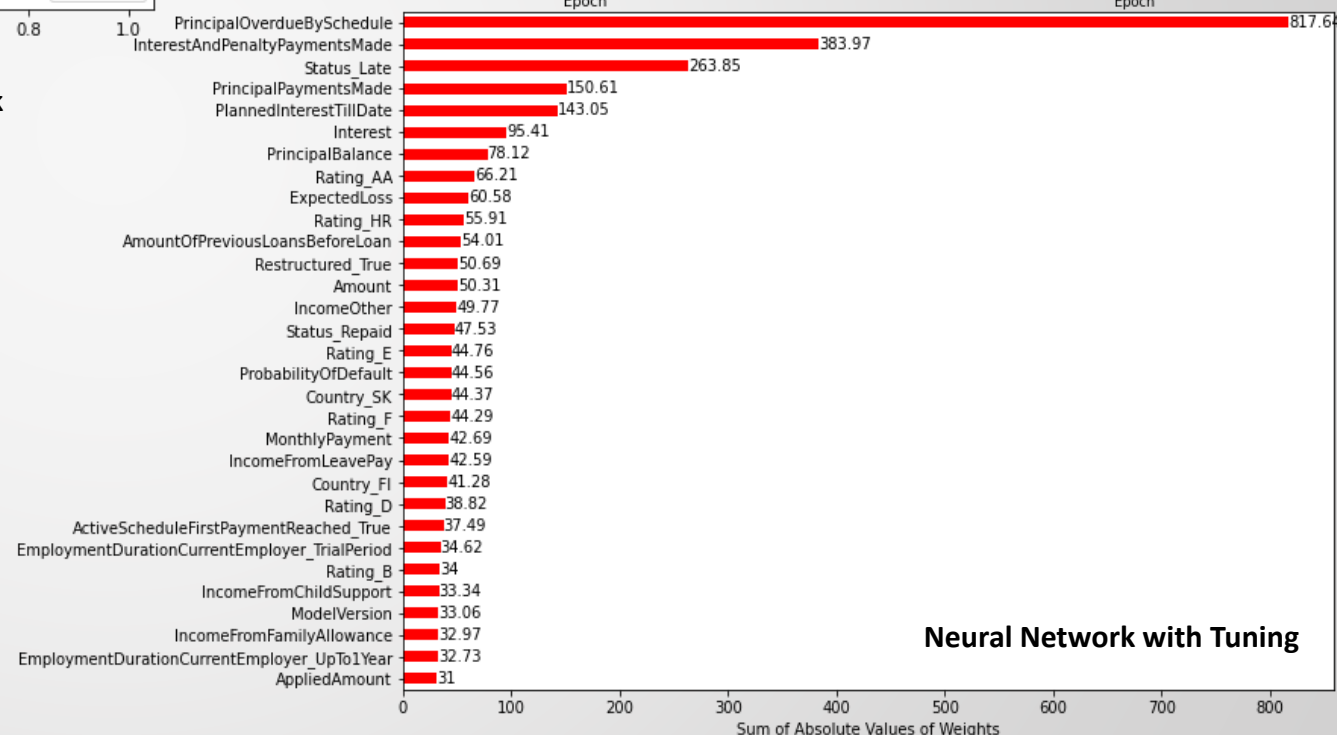
- ✓ 3 Hidden Layers: 100, 50, and 25 Neurons, Relu Activation
- ✓ 1 Output Layer, 1 Neuron, Sigmoid Activation
- ✓ Grid Search CV = 3
- ✓ Hyperparameters
 - Optimizer: rmsprop, adam
 - inits: glorot_uniform, normal, uniform
 - Epochs: 50,100
 - Batches: 5,20



Neural Network



Neural Net	Class 0 Predicted	Class 1 Predicted
Class 0 Actual	630	308
Class 1 Actual	44	3,018



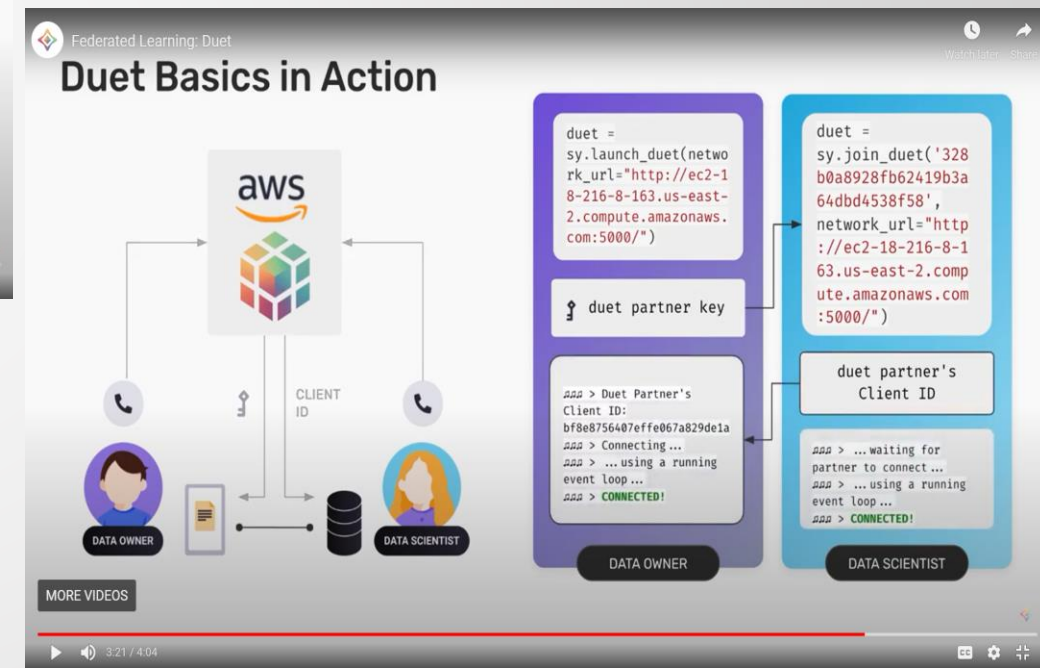
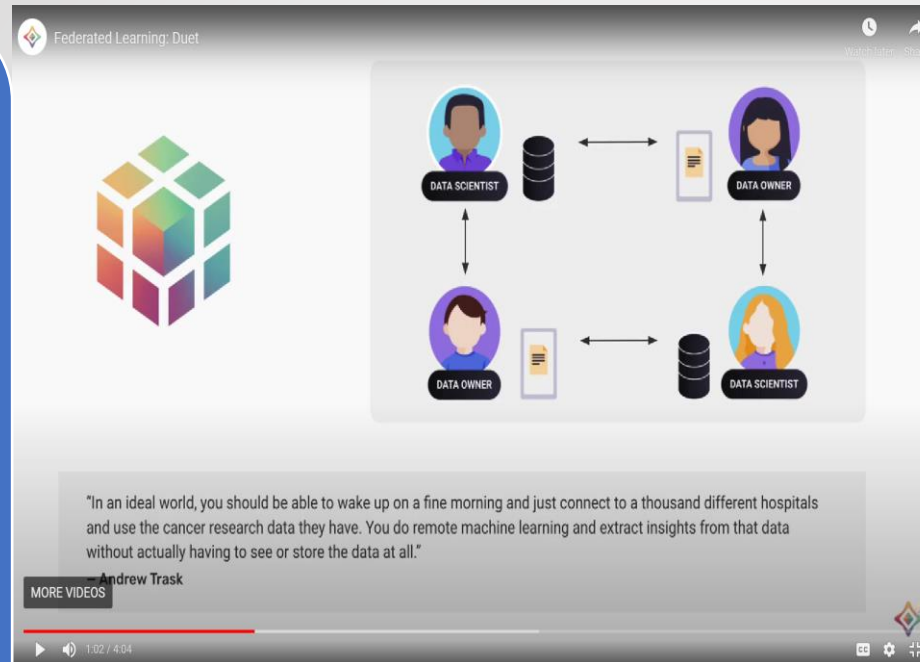
Remote Machine Learning - Overview

Why Useful?

- ✓ Keeps Data Private
- ✓ Data Owner has Control Over Data
- ✓ Machine Learner Benefits from Access to Distributed Data

Process?

- ✓ PySft Wrapper to ML Package
- ✓ Encryption and Privacy Maintained
- ✓ Machine Learner Can Access Multiple Data Sources Simultaneously
- ✓ Models Trained Remotely and can be Aggregated for Use



Remote Machine Learning –PyTorch/PySft Results

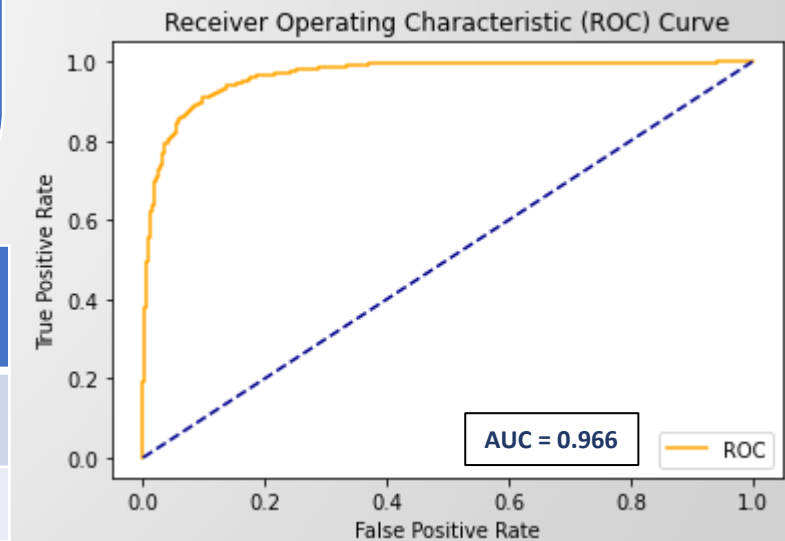
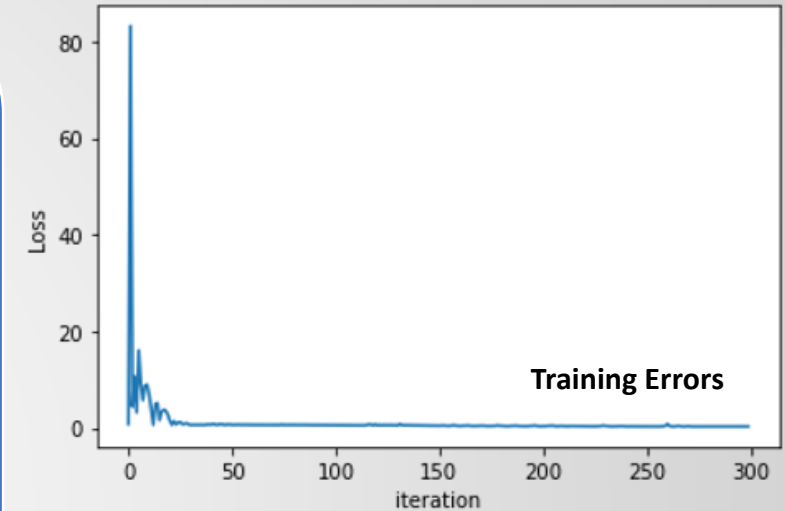
Remote Learning Process:

- ✓ Data Owner/Data Scientist interact via PySyft and PyGrid/AWS
- ✓ Data Owner sends data to Data Scientist
- ✓ Data Scientist makes requests via Pysft to Data Owner
- ✓ Data Scientist creates model
- ✓ Data Scientist sends model to Owner
- ✓ Training on Remote Server
- ✓ Model Sent to Data Scientist Once Trained
- ✓ Data Scientist Tests Model – Sckit Learn Packages

PyTorch and PySft:

- ✓ 2 Hidden Layers: 100 and 100 Neurons, Relu Activation
- ✓ 1 Output Layer, 2 Neurons, Log_soft_max Activation
- ✓ 300 Epochs
- ✓ Optimizer: Adam
- ✓ learning_rate = .01
- ✓ nn.functional.nll_loss

PyTorch/ PySft	Class 0 Predicted	Class 1 Predicted
Class 0 Actual	1,262	99
Class 1 Actual	95	634





Model Evaluation – Performance Metrics

	Hyperparameters	RMSE	Accuracy	Precision	Recall	F_1Score	AUC
Logistic Regression	L1 Penalty, liblinear Solver, C =5	0.209	0.956	0.938	0.936	0.937	0.951
Naïve Bayes	Alpha = 1.0	0.399	0.841	0.789	0.743	0.765	0.818
Decision Tree	Criterion – entropy, Max_depth = 20	0.166	0.973	0.962	0.960	0.961	0.970
Ensemble Forest	N_estimators= 100 l_rate = 1.0	0.231	0.947	0.934	0.913	0.923	0.939
Random Forest	N_estimators = 200, Criterion – entropy, Max_features = auto	0.166	0.972	0.976	0.943	0.960	0.966
Neural Net – Keras/Tensorflow	Batch_size = 5, epochs=150, init- glorot_uniform, optimizer= adam	0.249	0.912	0.907	0.986	0.945	0.980
Neural Net - PyTorch	To be Developed		0.907	0.865	0.869	0.867	0.966

CONCLUSIONS - FORTHCOMING

Next Steps, To be Developed

- What is the significance of your project?
- How are the stakeholders affected by the outcome of your project?
- What are your recommendations for improvements for researchers who may continue your work?
- How will you build on this project in your work or academic career (optional)?

Questions

