Predicting Loan Approval

Courtney Mazzulla, Dan Nealon, Karsyn Lee, Sunny Shin

April 18, 2023

https://github.com/sunnyshin0824/207_final_project

Agenda

1 Motivation

Introduce the topic and its application to real world problems

2 Data

Describe the source of the dataset and pre-processing

3 Approach

Highlight the baseline model and additional improvements

4 Experiment

Describe different hyperparameter choices to improve the performance

5 Conclusion

Summarize the key results of the analysis and propose future work

Question: *Can we predict which loans will get approved?*

Relevance:

In 2023, only 32% of banks in the US use artificial intelligence such as predictive analytics, speech recognition, and other models, to get a competitive edge in the market*

Unique Value:

We provide an affordable model for loan prediction (that determines who is eligible to receive a loan) that <68% of banks in America could benefit from.

*National Business Research Institute

Why would a bank want to use ML for loan prediction?

Optimize Time

Takes an average of 52 days for a bank to grant a loan, this model predicts instantly.

Reduce Bias

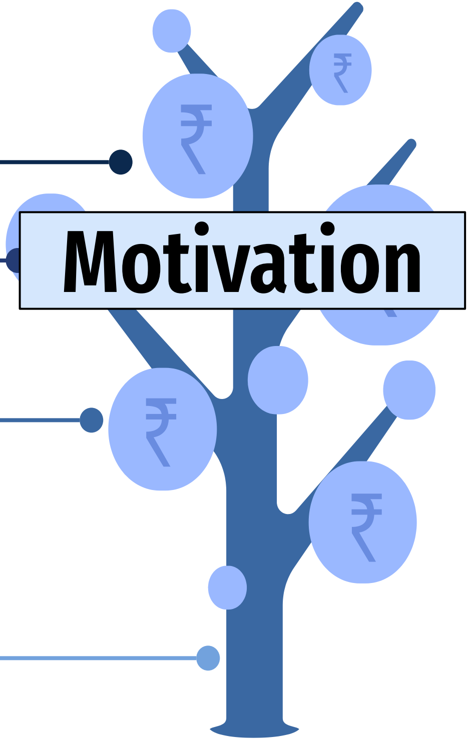
Reverse past and present discrimination in lending, and to foster a more inclusive economy by eliminating human biases.

Increase Accuracy

Banks are less likely to give loans to defaulters due to no human error & a model with increased accuracy.

Maximize Revenue

For banking companies that use human manual processing to distribute loans, this is a cost efficient way to reduce/optimize workforce and resources.



1 **Title:** Loan Prediction Problem Dataset

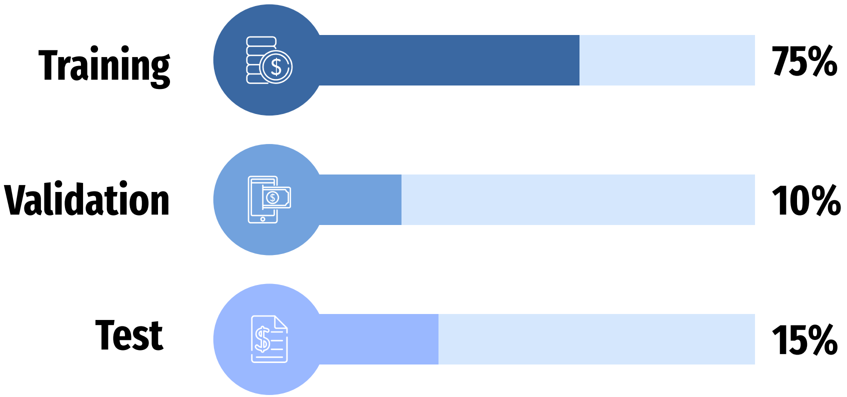
2 **Source:** Kaggle

3 **Size:** 12 x 615 training dataset

- Training dataset
- Validation dataset
- Test dataset

Statistical Summary of Dataset

	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History
count	614.000000	614.000000	592.000000	600.00000	564.000000
mean	5403.459283	1621.245798	146.412162	342.00000	0.842199
std	6109.041673	2926.248369	85.587325	65.12041	0.364878
min	150.000000	0.000000	9.000000	12.00000	0.000000
25%	2877.500000	0.000000	100.000000	360.00000	1.000000
50%	3812.500000	1188.500000	128.000000	360.00000	1.000000
75%	5795.000000	2297.250000	168.000000	360.00000	1.000000
max	81000.000000	41667.000000	700.000000	480.00000	1.000000



Data Types

1

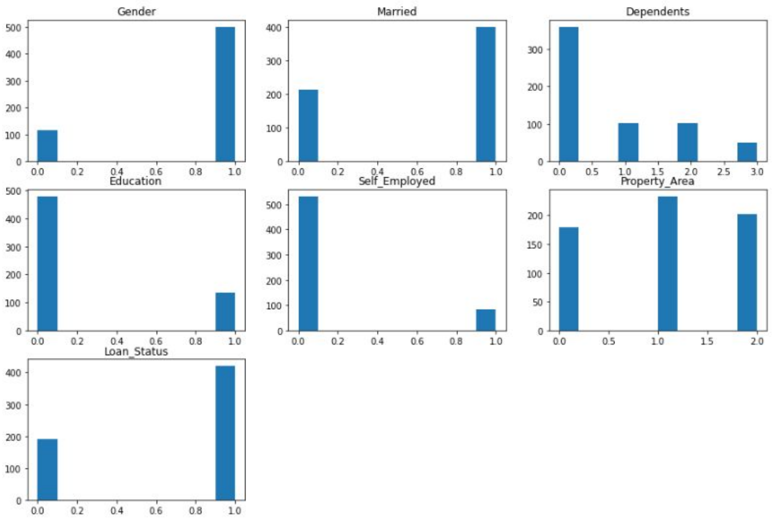
Categorical

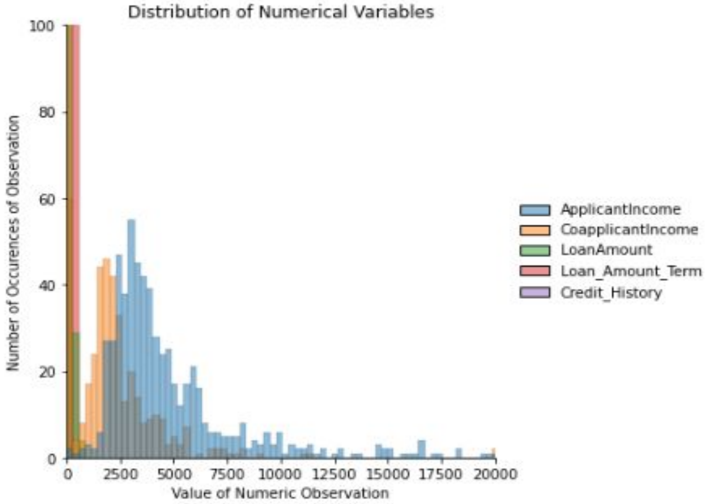
Variables including dependents, gender, marital status, education, self-employed, property area, loan_status

2

Numerical

Variables including applicant income, co-applicant income, loan amount, loan term, credit history

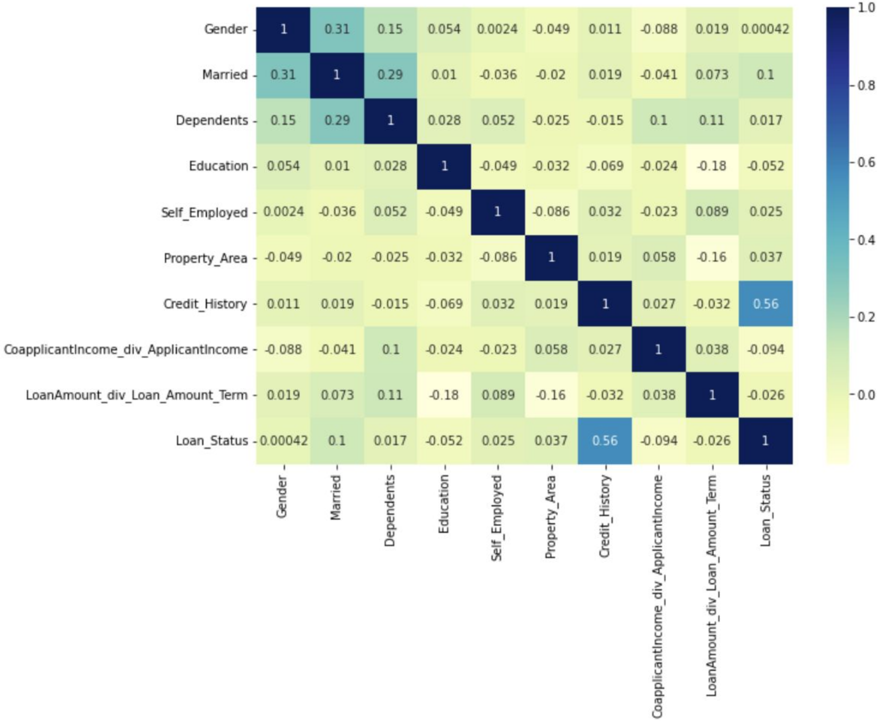




Data Preprocessing

- **Imputation**
  - Convert categorical into numerical values
  - Median Imputation for numerical
  - Mode Imputation for categorical
  - Conditional for gender on married
  - CART for Credit_History
- **Dimensionality Reduction**
  - Correlation-based Feature Selection
  - Combine features
- **Standardization**
  - StandardScaler()

Correlation Matrix Post-Data Manipulation



Baseline model + Improvements

Baseline Model

Model 1

Simple Logistic Regression

Model 2

Implement FFNN model

Model 3

Tune the hyperparameters

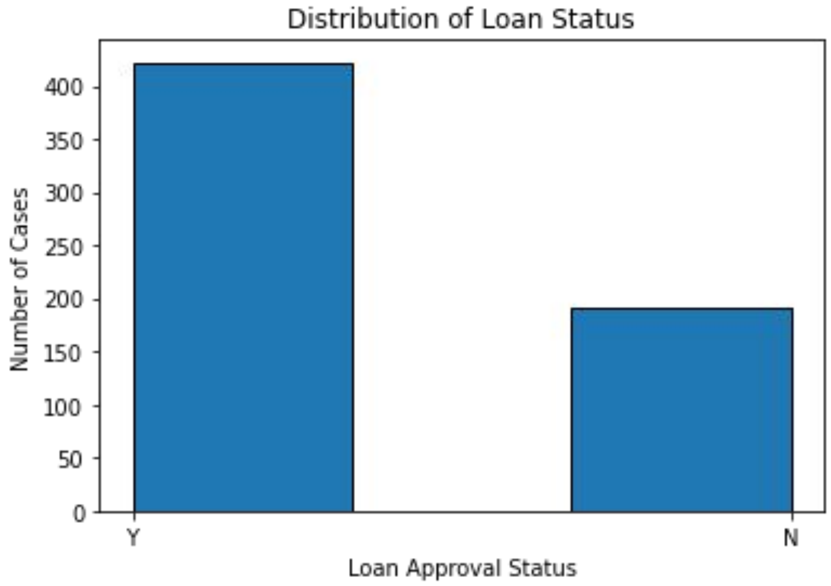
Model 4

Baseline Model

Considering the distribution of loan approvals in our training data, we decided to have our baseline model always predict approval.

This resulted in 70% accuracy on our training predictions.

Our next models would attempt to beat this performance.



Logistic Regression

Considering the classification requirement for any model we would construct to predict loan approvals, logistic regression would be a great first model.

Using Gender, Married, Dependents, Education, Self_Employed, Property_Area, Credit_History, Co-applicant Income/Applicant Income, LoanAmount / Loan_Amount_Term as predictors and loan status as our response, we were able to achieve the following performance with an 75/10/15 split of the data:

```
LogisticRegression :
Precision score: 0.816
Recall score: 0.977
F1 score: 0.889
Loss: 5.913
Accuracy: 0.829
```

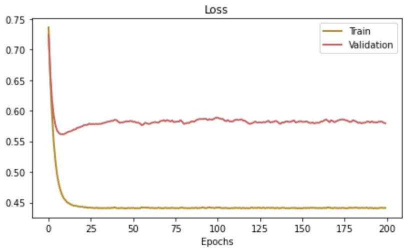
Precision is important as we don't want to approve a loan if we shouldn't (FP)

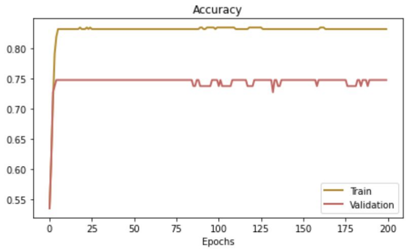
Accuracy improvement on baseline

FFNN

In an effort to improve our model's performance, we decided to build a neural network model with one dense layer and a sigmoid activation function, which makes it a binary classification model.

Our hope is to split the problem of classification into a layered network of simpler elements, and perform faster predictions long-term.






```
FFNN train accuracy: 0.834
FFNN val accuracy: 0.747
```

```
FFNN test accuracy: 0.805
```

Hyperparameter Tuning

Learning Rate

[0.001, 0.01, 0.1]

Optimizer

[Adam, SGD]

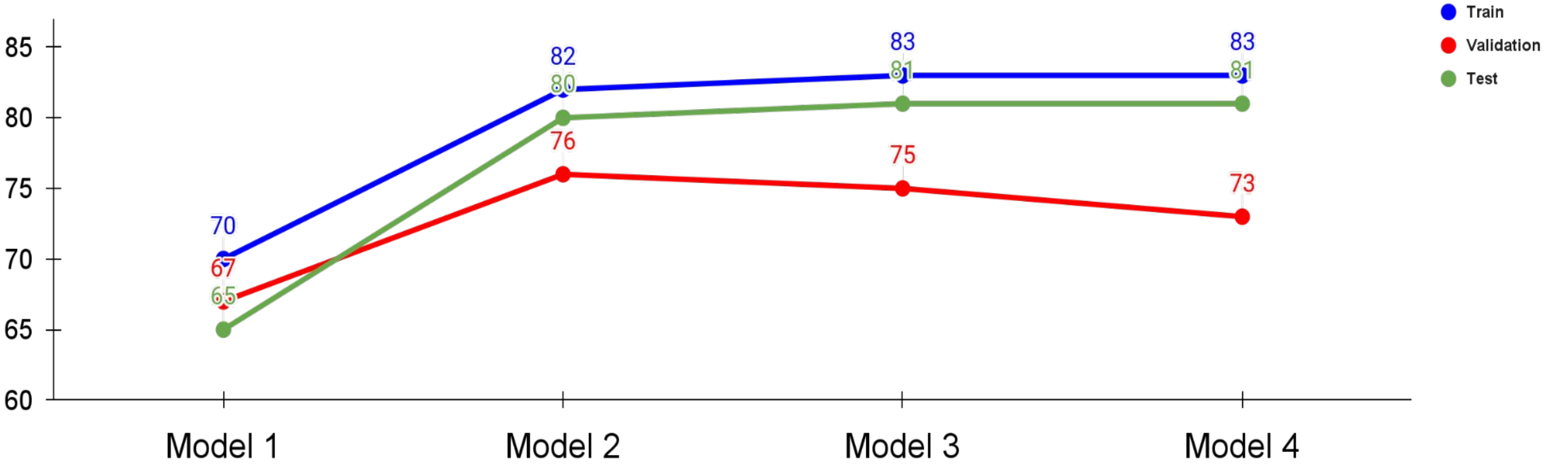
Epochs

[100, 200, 300]

	Learning Rate	Optimizer	Epoch	Training Accuracy	Validation Accuracy
0	0.001	Adam	100	0.831633	0.747475
1	0.001	Adam	200	0.831633	0.747475
2	0.001	Adam	300	0.831633	0.747475
3	0.001	SGD	100	0.655612	0.595960
4	0.001	SGD	200	0.813776	0.737374
5	0.001	SGD	300	0.829082	0.747475
6	0.010	Adam	100	0.834184	0.747475
7	0.010	Adam	200	0.831633	0.747475
8	0.010	Adam	300	0.831633	0.747475
9	0.010	SGD	100	0.829082	0.747475
10	0.010	SGD	200	0.831633	0.747475
11	0.010	SGD	300	0.831633	0.747475
12	0.100	Adam	100	0.836735	0.727273
13	0.100	Adam	200	0.829082	0.737374
14	0.100	Adam	300	0.823980	0.727273
15	0.100	SGD	100	0.831633	0.747475
16	0.100	SGD	200	0.831633	0.747475
17	0.100	SGD	300	0.831633	0.747475

```
Improved FFNN test accuracy: 0.8130
```

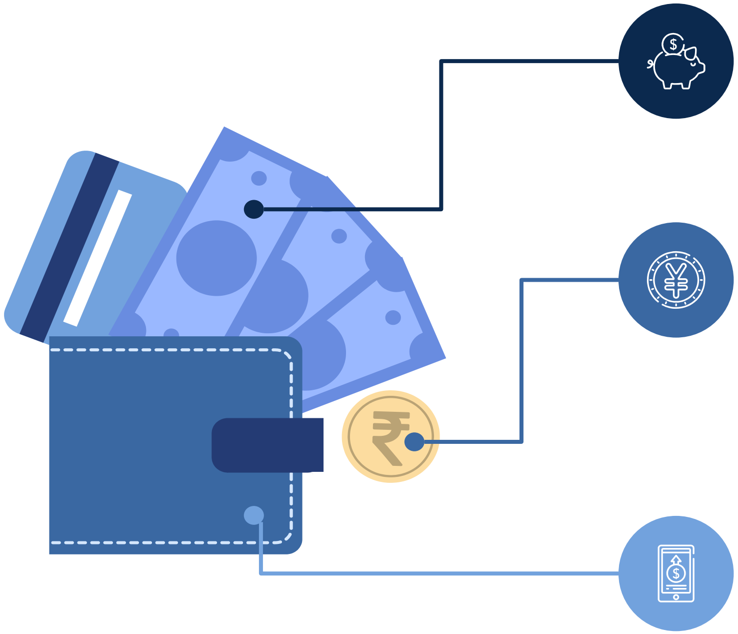
Performance Improvement



Model 1	Model 2	Model 3	Model 4
Baseline model	Logistic regression model	FFNN	FFNN with improved hyperparameters

Conclusion/Future State



Ensemble modeling

By combining the predictions of several models, using the ensemble modeling, we believe we can reduce the impact of individual model biases and errors, and produce a more accurate prediction.

Feature engineering

We can further improve the model by creating new features that capture additional information about the loan applicants. This can include socio-economic factors, geographic data, and other variables that may impact a borrower's likelihood of defaulting on a loan.

Incorporate new data sources

By incorporating new data sources such as credit scores, financial statements, and social media data, we can increase the accuracy of the model and make it more practical/user friendly in the market.

Thank You!

Team Contribution

	Courtney Mazzulla	Dan Nealon	Karsyn Lee	Sunny Shin
Data Research	X	X	X	X
Environment Set-up	X	X	X	X
Data Cleaning	X	X	X	X
Data Splitting	X	X	X	X
Hyperparameter tuning	X	X	X	X
Presentation slides	X	X	X	X