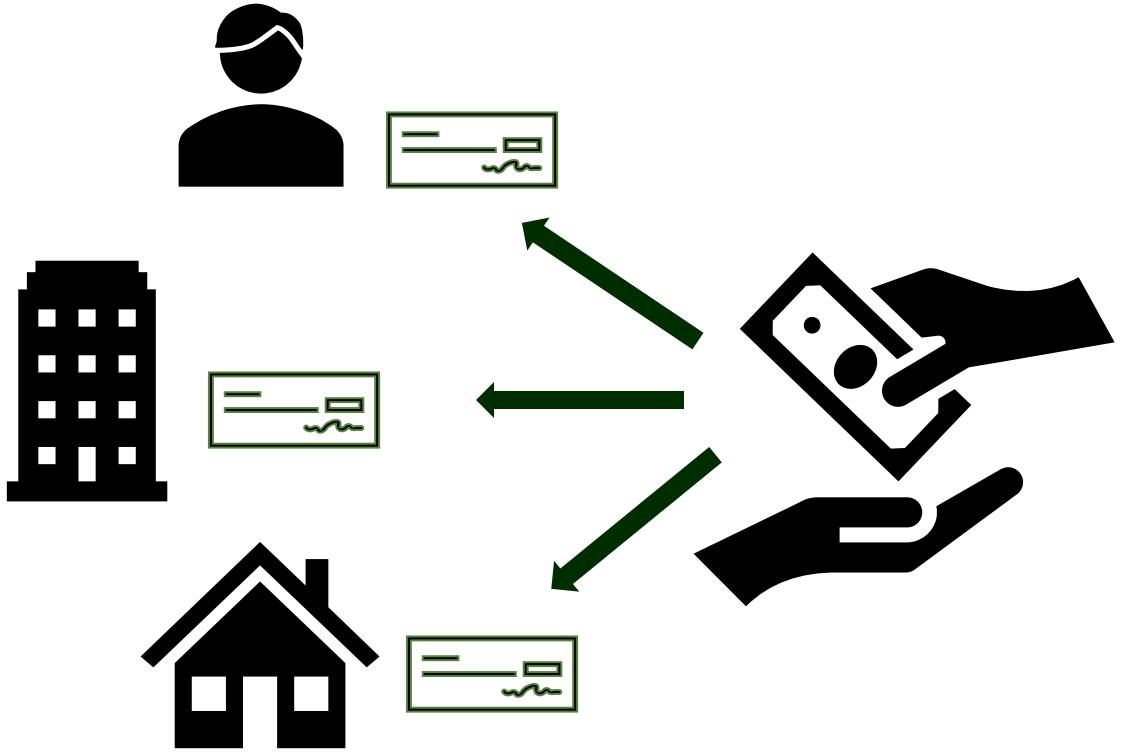
Intro to Data Science – Final Group Project

Team Name: Lisbon
Team Members: Jessica Roman, Austin Zhang, Hana Ghattas, Jonah Crystal, Quinn Kerrigan, Tina Xu
Dec 4, 2023

Understanding Debt

Financial Institutions lend money in different forms to people, governments, organizations, other banks, etc. In exchange for the loan, they receive interest payments and are eventually paid back for the full amount of the loan, called the principal amount

Borrowers



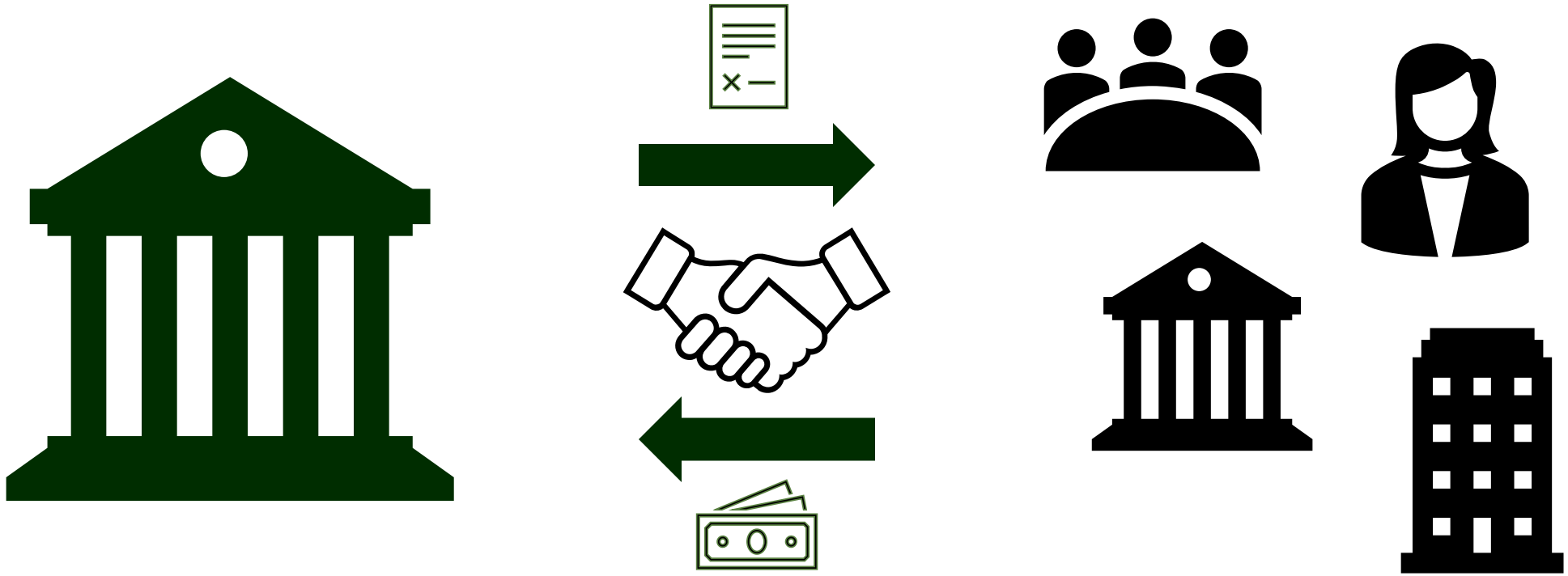
Financial Institutions

Understanding Fixed Income

Fixed income is a type of investment that allows investors to lend money to institutions, in exchange for a promise to pay them back. In addition, the investors receive a flow of fixed payments

Financial Institutions

Investors / Lenders



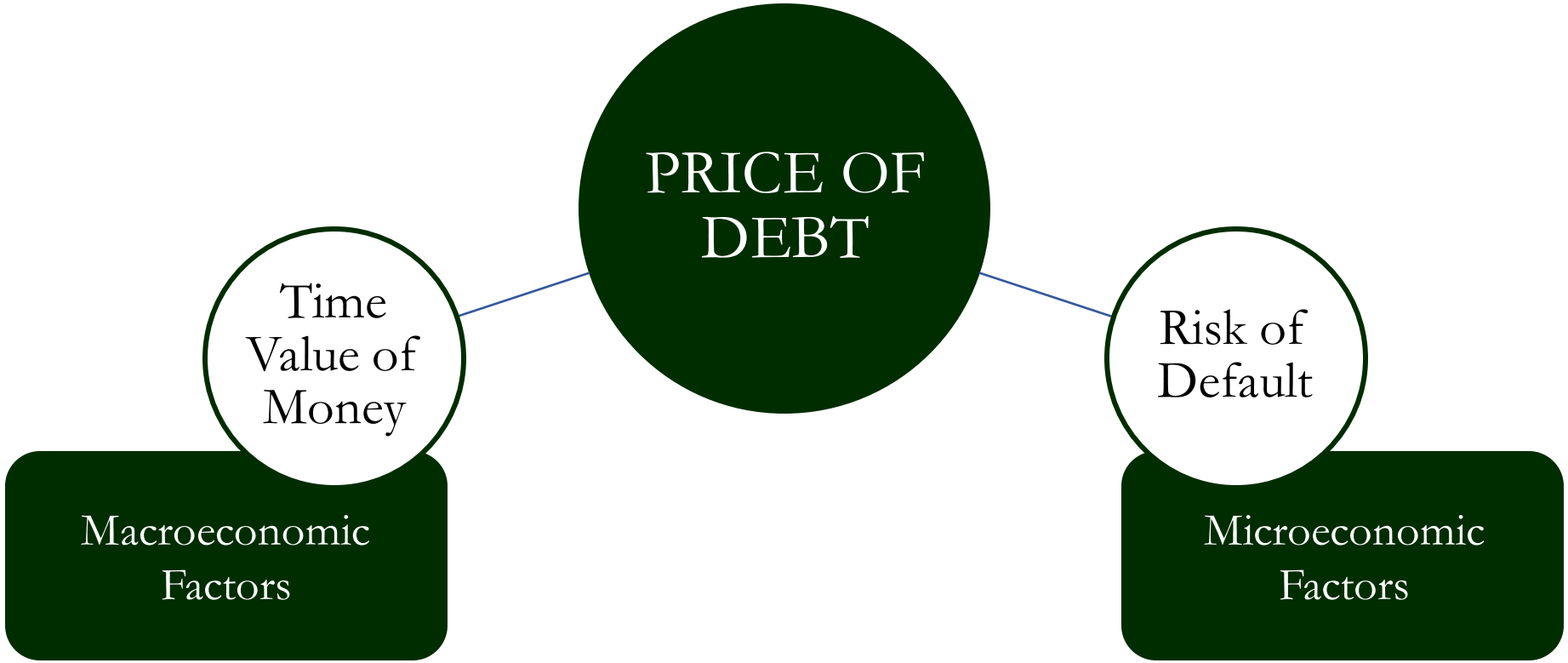
Understanding Fixed Income

Putting these together, banks often package a variety of the loans that they issue into securities that are bought by investors. In this case, the fixed payments are made up of the interest payments on the debt

Borrowers	Financial Institutions	Investors / Lenders

Credit Risk

The debt is not always paid back. Sometimes individuals or organizations go bankrupt are not able to repay their debt. This is called default



Who Cares

Investors must accurately predict the probability of default in order to accurately price the loan

Accurately Predict Defaults

- Price the security accurately
- Seek out mis-priced debt
- Less risk, more reward

How to Assess Risk

Various factors are weighed to determine the likelihood of default, and in turn, to assess creditworthiness and set interest rates

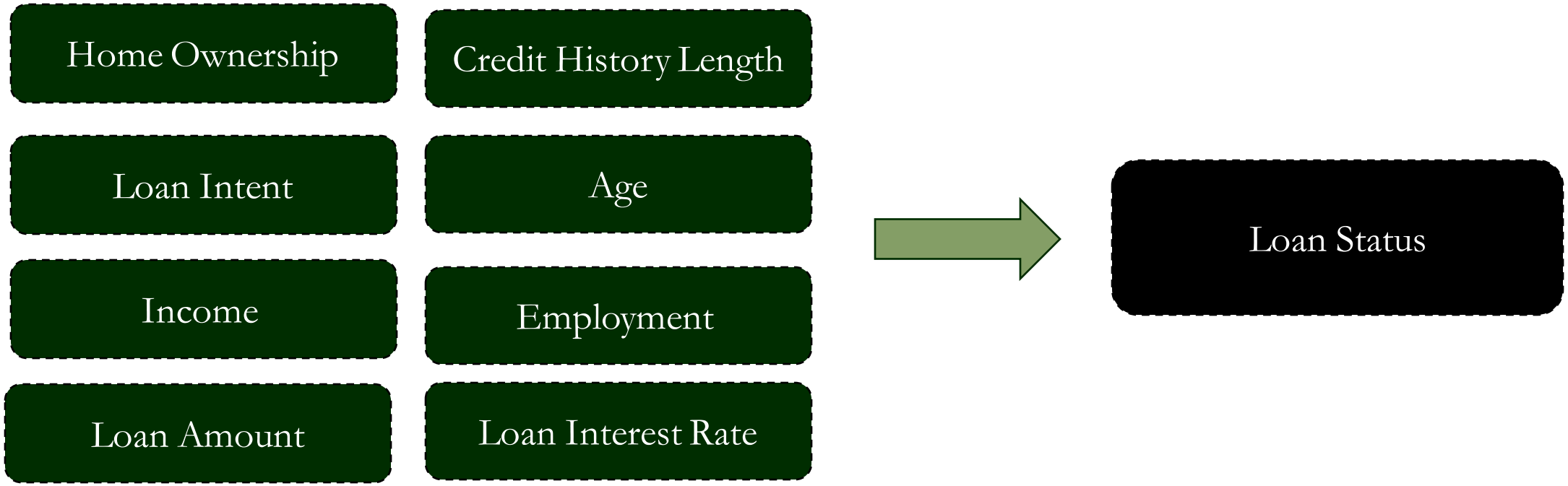
Data Analysis

The goal of the model is to predict if a person will ***default on a loan,*** *so that* ***lenders*** *can find debt that is trading at a discount*

- Home Ownership
- Credit History Length
- Loan Intent
- Age
- Income
- Employment
- Loan Amount
- Loan Interest Rate

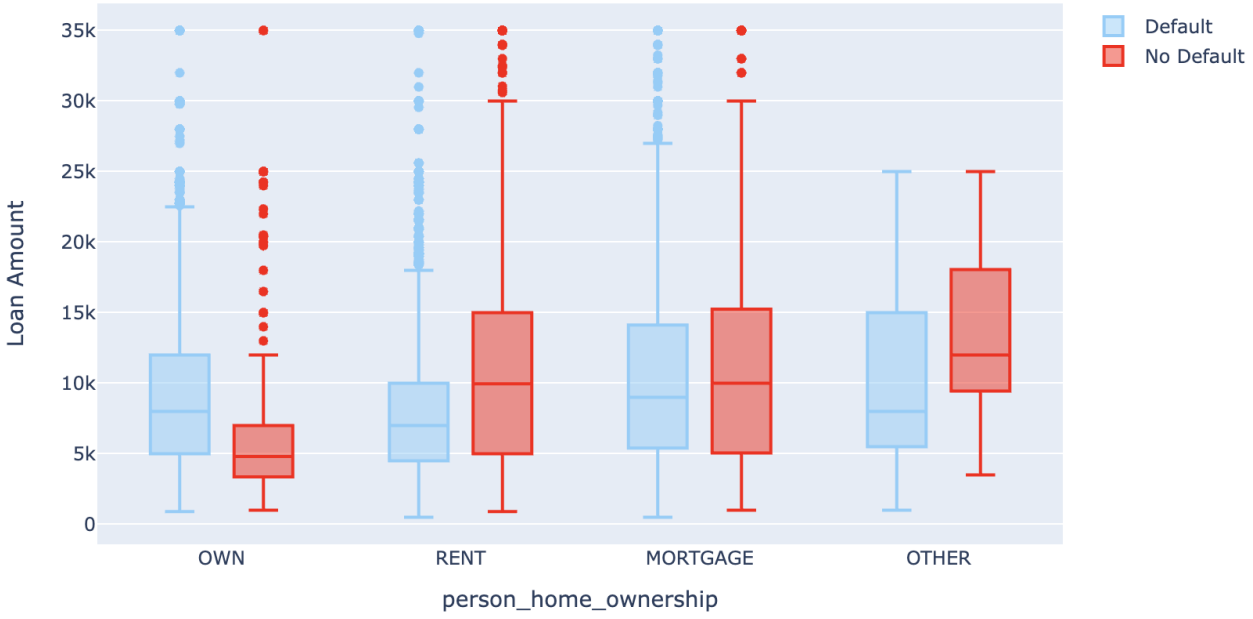
Data Analysis

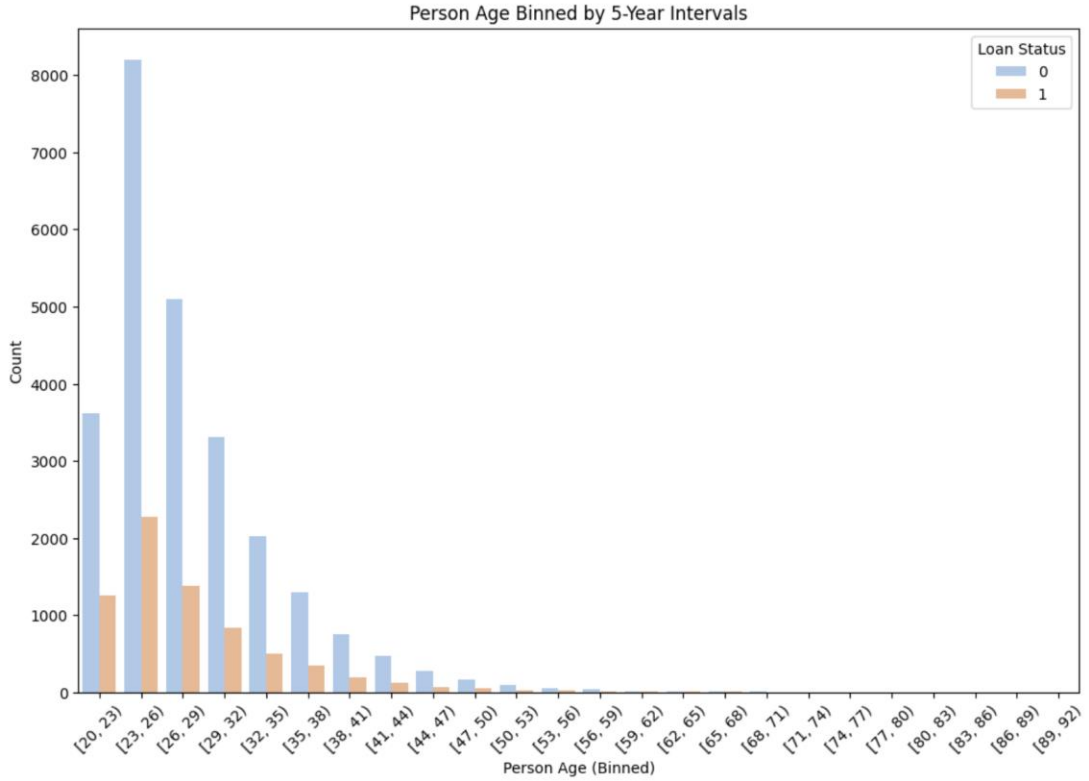
The goal of the model is to predict if a person will ***default on a loan,*** *so that* ***lenders*** *can find debt that is trading at a discount*



Data Analysis

People with homes generally have better credit & sampling bias in the data





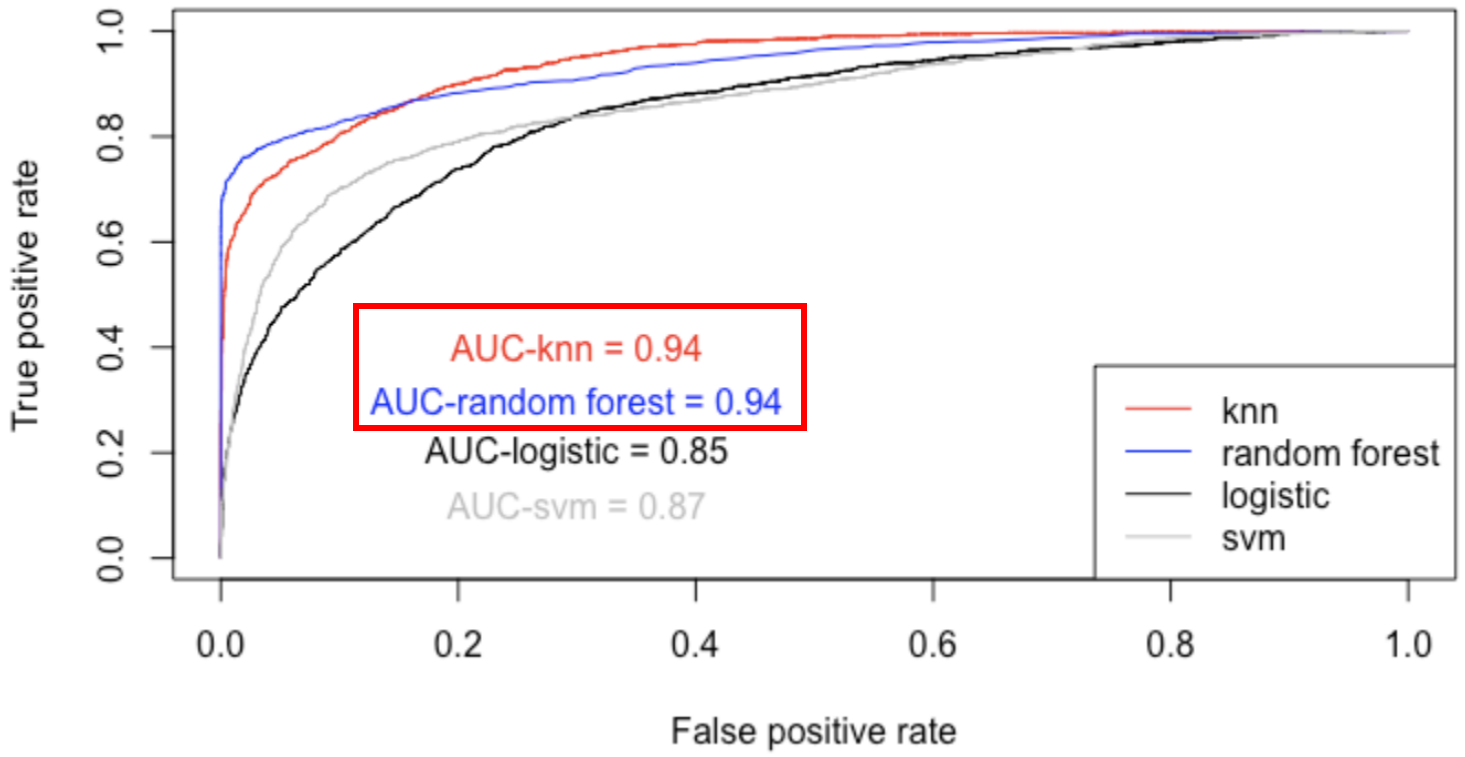
Data Analysis

We considered four models

- Linear Regression
- KNN
- Random Forest
- Logistic Regression
- Support Vector Machines

Assessment

Each model's result plotted in the ROC graph and evaluated with Area Under Curve



Assessment

Which one fits our criteria better?

Random Forest	Reference	
Prediction	0	1
0	4471	347
1	26	884
Accuracy: 0.9349		

FNR: 0.28

KNN	Reference	
Prediction	0	1
0	4405	468
1	57	798
Accuracy: 0.9083		

FNR: 0.37

Key Criteria:

High Model Accuracy

Low False Negative Rate

Assessment

Random Forest is the final winner that beats the two criteria

Random Forest	Reference	
Prediction	0	1
0	4471	347
1	26	884
Accuracy: 0.9349		

KNN	Reference	
Prediction	0	1
0	4405	468
1	57	798
Accuracy: 0.9083		

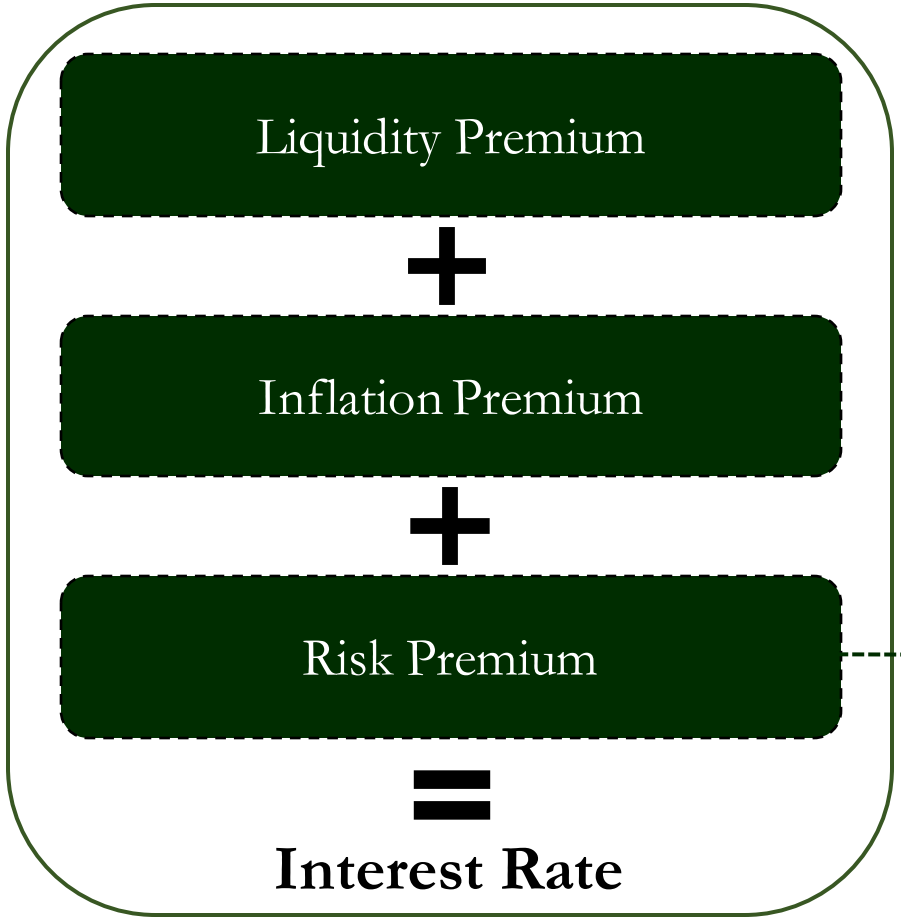
Key Criteria:

High Model Accuracy

Low False Negative Rate

Implementation

Investors could use this model to appropriately price debt / packages of debt



By using the model to determine default risk, investors could establish the appropriate risk premium, and calculate an adequate interest rate

Ethical Implications

Both data privacy and output fairness are crucial part of ethical implication of the model

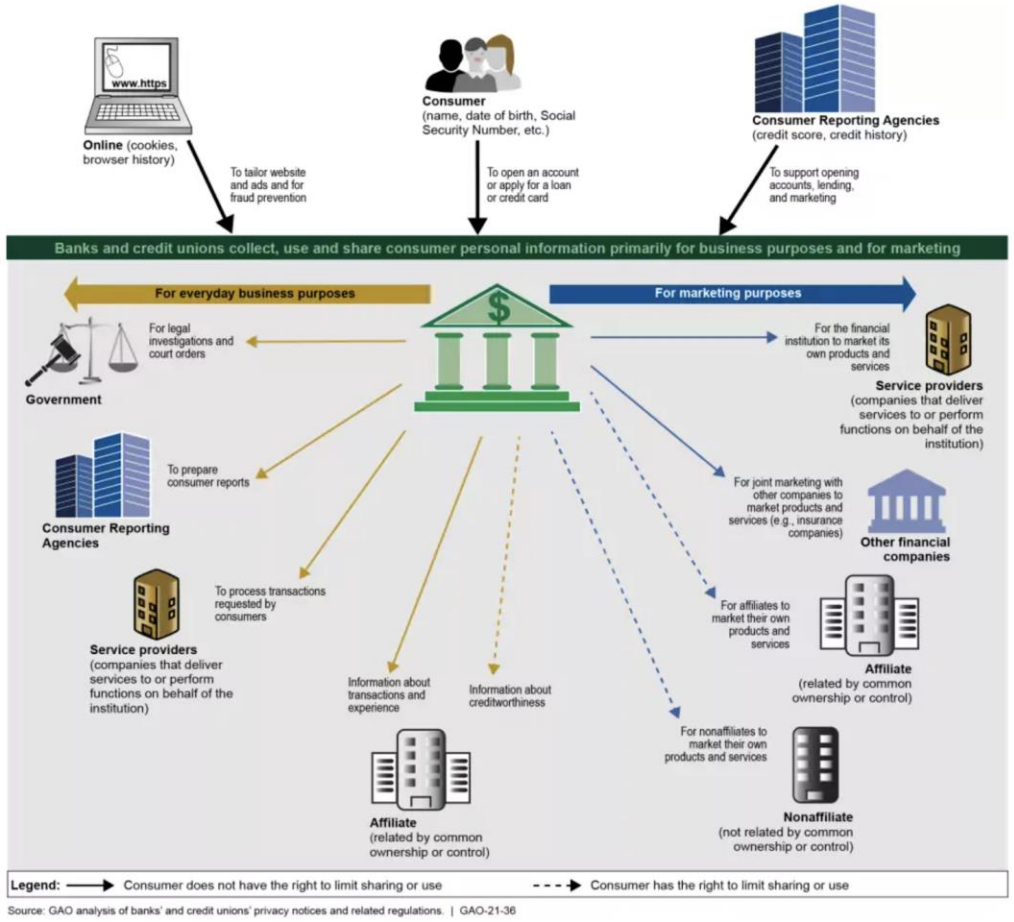
Variables

- Home Ownership
- Credit History Length
- Loan Intent
- Age
- Income
- Employment
- Loan Amount
- Loan Interest Rate

Output Predictive Model

Ethical Implications

Both Data Privacy and Output Fairness are crucial part of Ethical Implication of the model



Appendix

Data Analysis

The goal of the model is to predict if a person will ***default on a loan,*** *so that* ***lenders*** *can find debt that is trading at a discount*

	person_age <int>	person_income <int>	person_home_ownership <chr>
1	22	59000	RENT
2	21	9600	OWN
3	25	9600	MORTGAGE
4	23	65500	RENT
5	24	54400	RENT
6	21	9900	OWN

person_emp_length <dbl>	loan_intent <chr>	loan_grade <chr>	loan_amnt <int>	loan_int_rate <dbl>
123	PERSONAL	D	35000	16.02
5	EDUCATION	B	1000	11.14
1	MEDICAL	C	5500	12.87
4	MEDICAL	C	35000	15.23
8	MEDICAL	C	35000	14.27
2	VENTURE	A	2500	7.14

loan_percent_income <dbl>	cb_person_default_on_file <fctr>	cb_person_cred_hist_length <int>
0.59	1	3
0.10	0	2
0.57	0	3
0.53	0	2
0.55	1	4
0.25	0	2

loan_status <fctr>
1
0
1
1
1
1

The structure of the data:

Cols:

person_age (NUMERIC)
person_income (NUMERIC)
person_home_ownership (FACTOR)
Person_emp_length (NUMERIC)
loan_intent (FACTOR)
loan_grade (FACTOR)
loan_amnt (NUMERIC)
loan_int_rate (NUMERIC)
loan_status (0/1) = y
loan_percent_income (NUMERIC)
cb_person_default_on_file (0/1)
cb_person_cred_hist_length (NUMERIC)

Additional Relationships

Some other histograms regarding frequencies of different columns. (https://colab.research.google.com/drive/1DAUNbjThCl982-yP_dZO7vRI2YV2XQzZ?usp=sharing)