Loan Default Analysis

By: Le Cai, Scott Chang, Henry Collins, Quincy Kay, Jack Van Zeeland

General Overview
+
Introductory Graphs

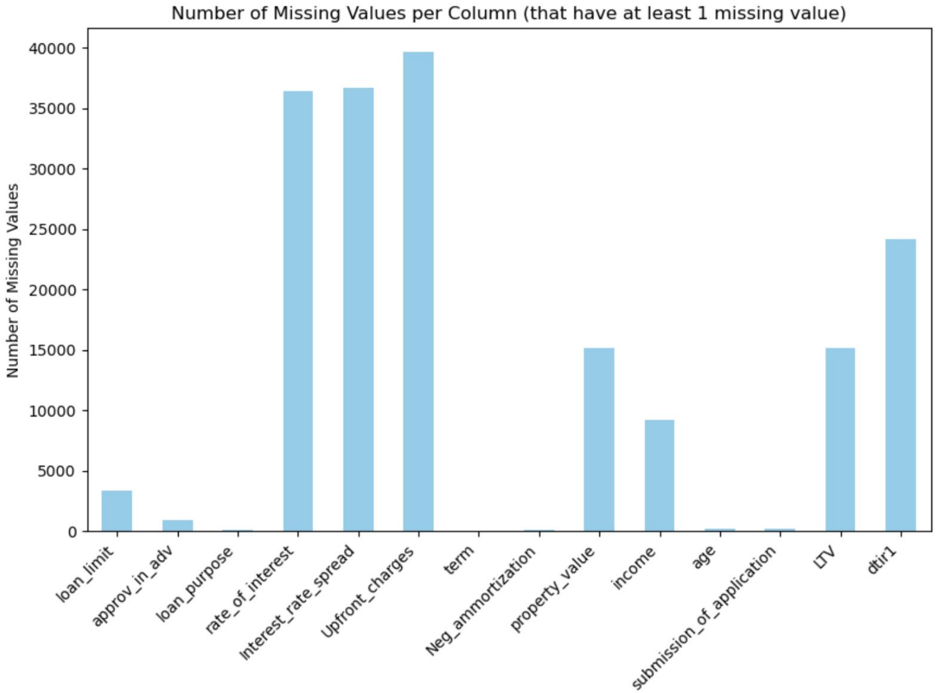
Overview:

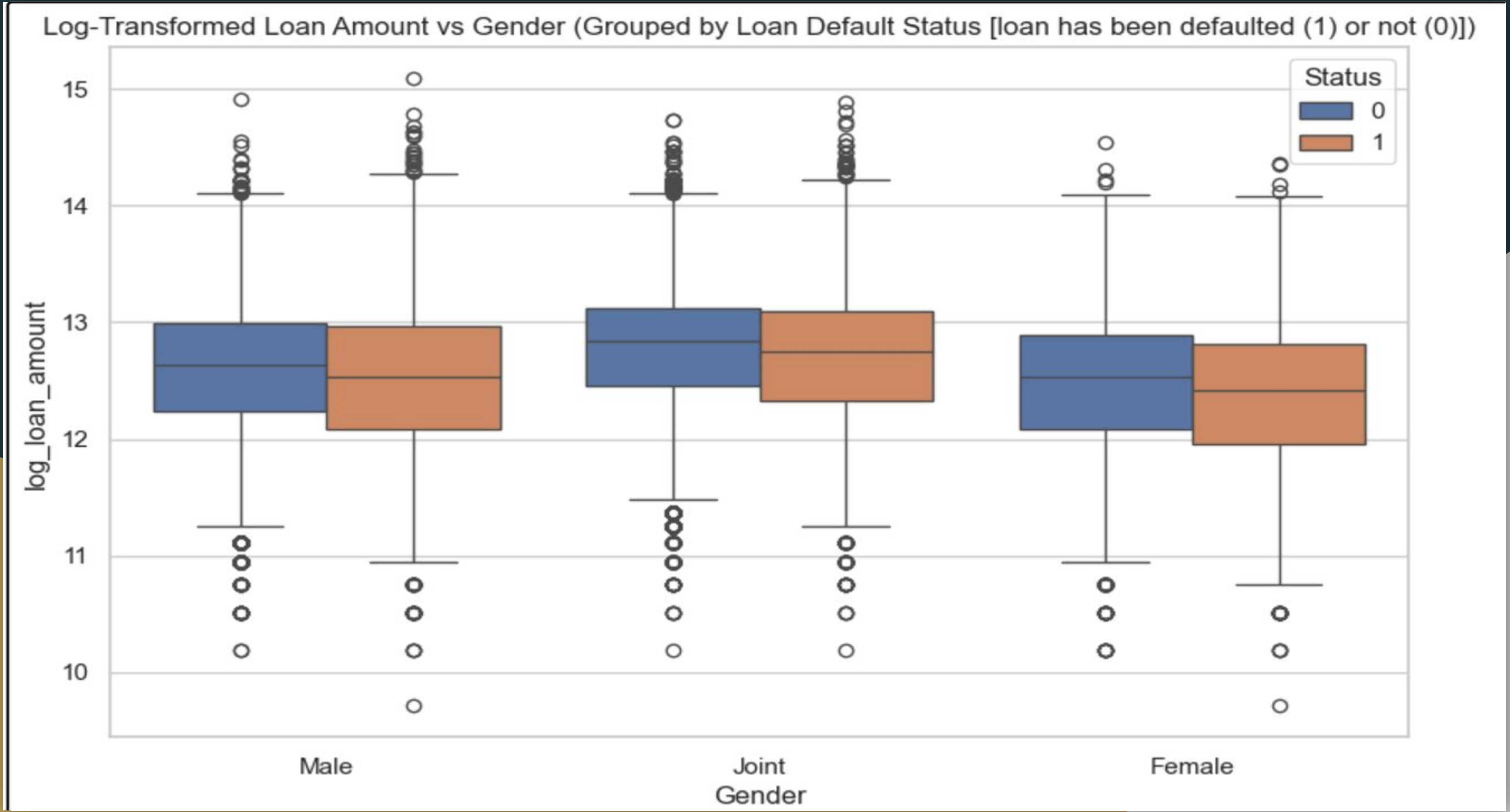
- Data - Kaggle Loan Default Dataset
- Size - 149K data points
- Influence Factors - Gender, loan type, loan purpose, income, region, age, etc.
- Outcome Variables - loan amount, interest rate, default status, etc.

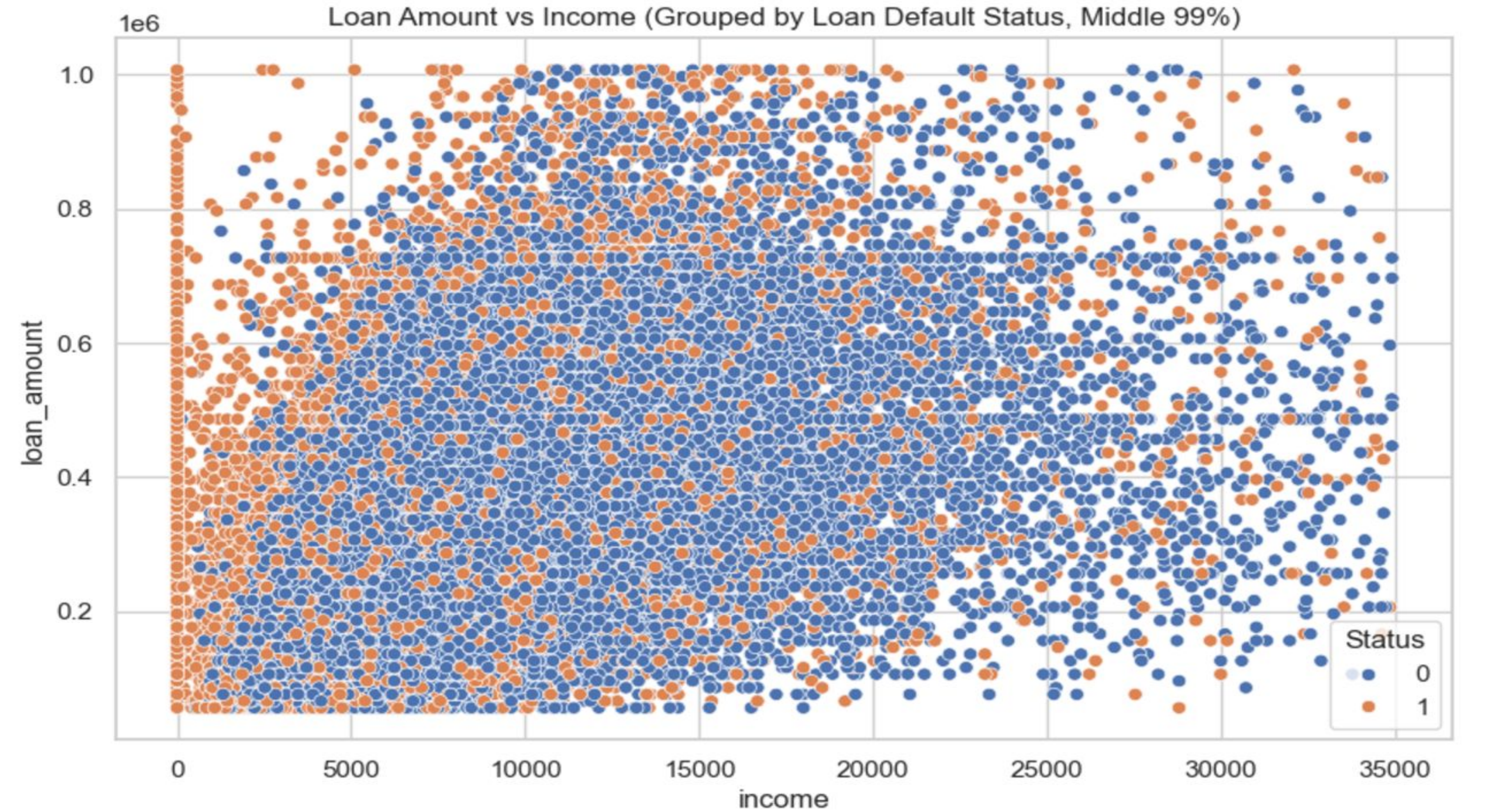
Data Clean-Up:

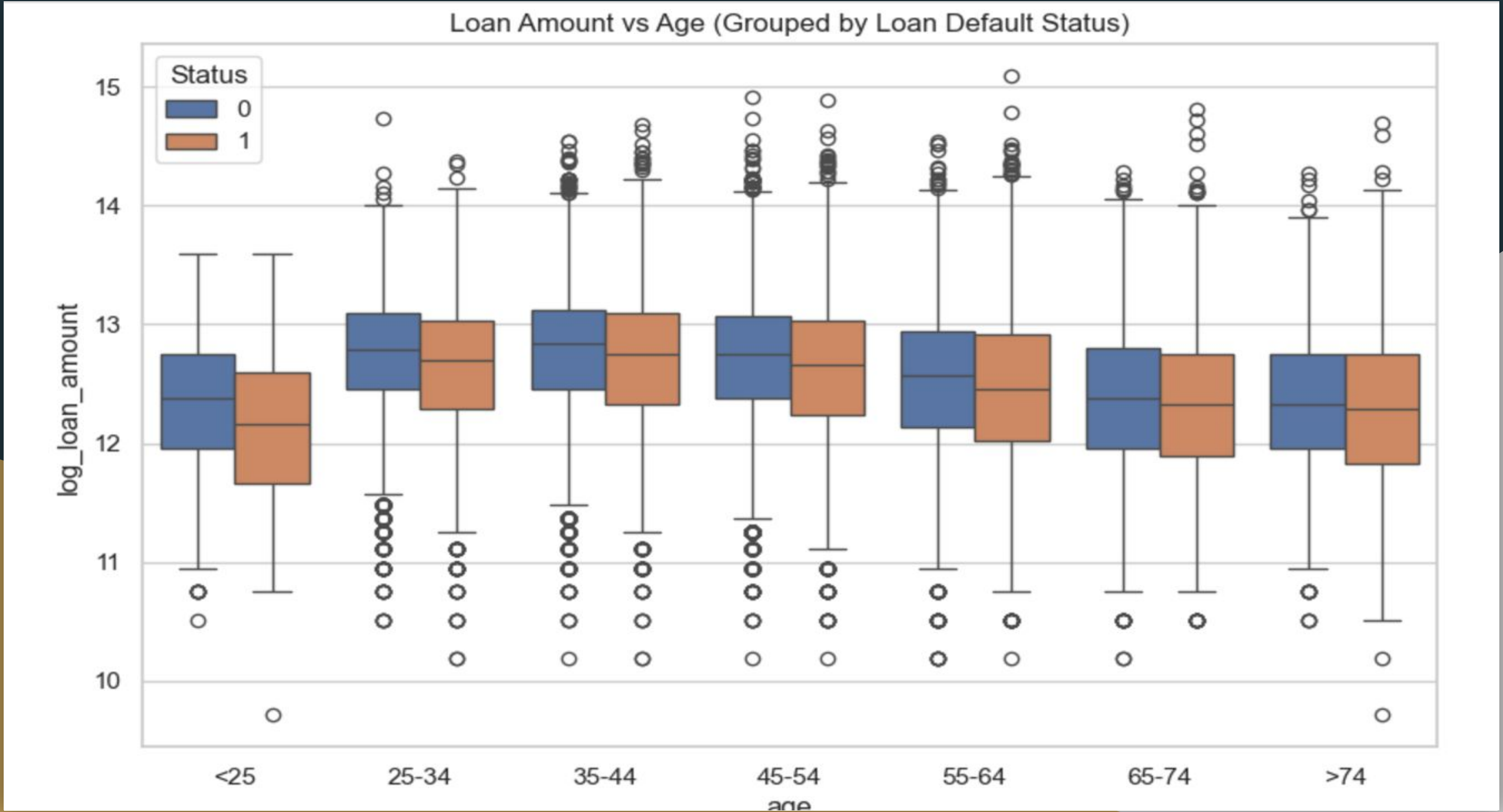
- 21 categorical columns
  - One-hot encoding
    - Used drop_column = True
      - Prevents collinearity

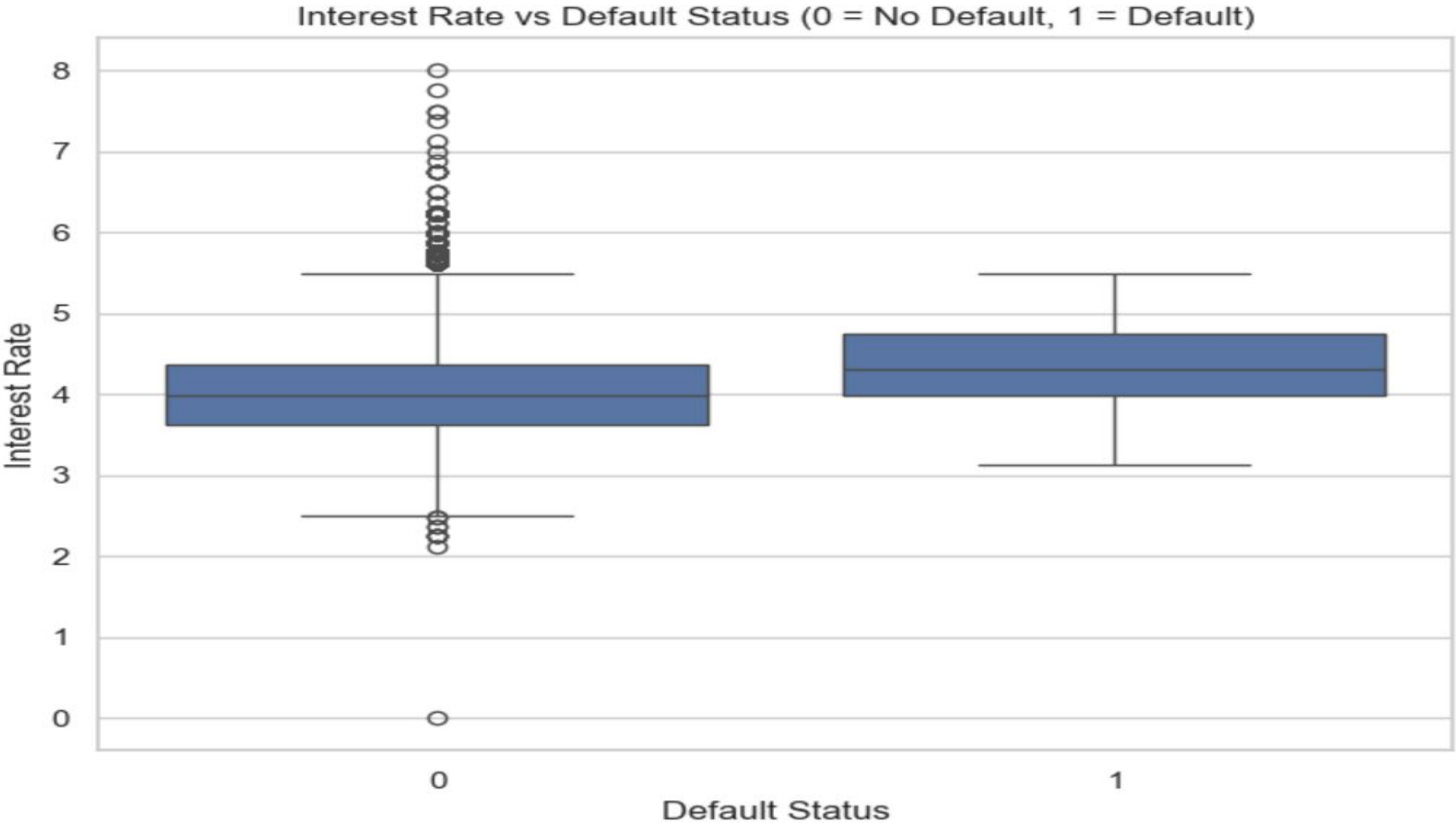
- 31% of rows have missing values
    - Drop or data imputation approach?
      - Addressed in our questions




Log-Transformed Loan Amount vs Gender (Grouped by Loan Default Status [loan has been defaulted (1) or not (0)])
Status
🟦 0
🟧 1
log_loan_amount
15
14
13
12
11
10
Male
Joint
Female
Gender


Loan Amount vs Income (Grouped by Loan Default Status, Middle 99%)
1e6
1.0
0.8
0.6
0.4
0.2
loan_amount
0
5000
10000
15000
20000
25000
30000
35000
income
Status
0
1


Loan Amount vs Age (Grouped by Loan Default Status)
Status
🟦 0
🟧 1
log_loan_amount
15
14
13
12
11
10
<25
25-34
35-44
45-54
55-64
65-74
>74
age


Interest Rate vs Default Status (0 = No Default, 1 = Default)
Interest Rate
8
7
6
5
4
3
2
1
0
0
1
Default Status

Topic 1:

- What algorithm is best at classifying if a person will default?
- What does this algorithm tell us about important features in predicting default rates?

Topic 1:

What algorithm is best at classifying whether a person will default?

Approach to train and assess each model

1. Choose models to consider
2. Prepare the data
   a. Handle NaN Values
   b. Feature engineering
   c. Split data
3. Train models and optimize hyperparameters
   a. Grid search to 'optimize' hyperparameters - *what are we optimizing for?*
4. Identify best classifier

Topic 1:

What algorithm is best at classifying whether a person will default?

Step 1: Choose Models to Compare

Notably, a person defaulting is a binary classification, so we chose 3 classifier models to consider:

Model	Hyperparameters for Training Model
Logistic Regression	- C parameter (controls overfitting)
ID3 Decision Tree	- *d* depth of tree - ϵ entropy split threshold
kNN Classifier	- *n* neighbors - Distance type

Topic 1:

What algorithm is best at classifying whether a person will default?

Step 2: Prepare the data

a. Handling NaN values
   a. ~⅓ of the data is NaN, so decided not to drop rows with missing values
   b. Instead, replaced missing values with kNN 5 nearest neighbors data imputation

b. Feature engineering
   a. Imputed data frame has 48 columns!
   b. Computationally unrealistic to train models on all these categories, so used SelectKBest to choose 15 best scoring categories to train on

c. Split data
   a. Chose 80% training, 10% validation, 10% test split

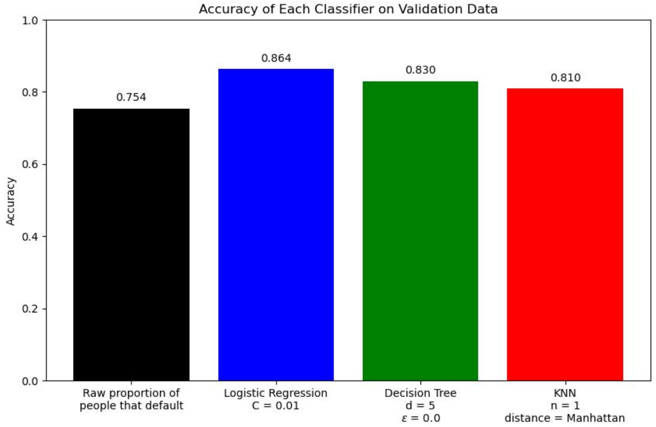
Topic 1:

What algorithm is best at classifying whether a person will default?

Step 3: Train models and optimize hyperparameters

- Grid search to optimize hyperparameters
- What are we optimizing for?
    - From the bank's perspective, we may want to maximize recall: Correctly identifying someone who actually does default (i.e. minimize false negatives)!
        - Used recall_scoring for GridSearch scoring

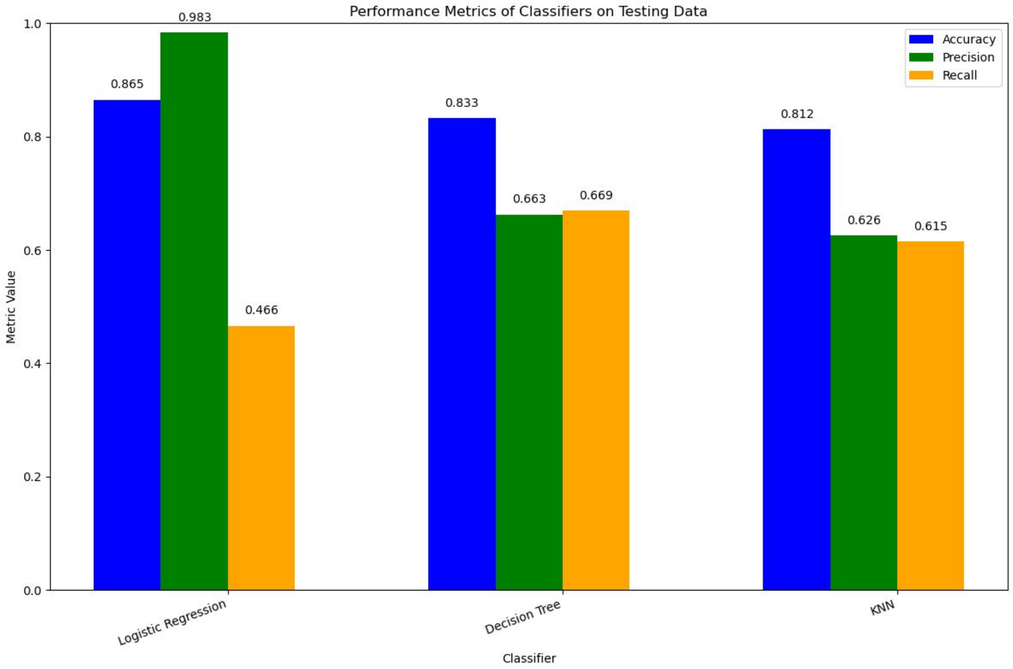
Parameter	**Parameter Values**
C (*LR*)	0.1, 1, 100, 1000
Depth (*DT*)	5, 50, 100
Entropy Split (*DT*)	0.0, 0.1
Neighbors (kNN)	1, 3, 5, 7
Distance type (kNN)	Euclidian, Manhattan



Topic 1:

What algorithm is best at classifying whether a person will default?

Step 4: Identify best classifier



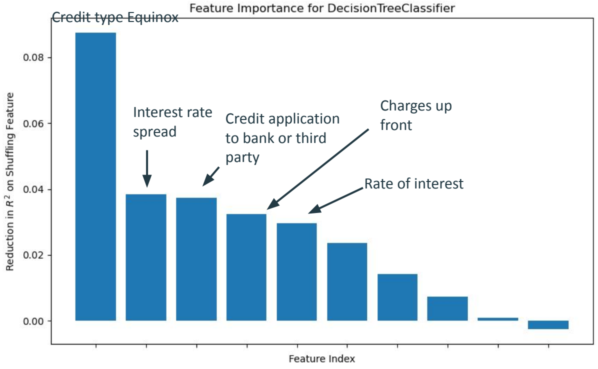
Takeaways:

- **Linear regression is most precise but worst at recall**
  - Out of all the predicted defaulters, most of them do actually default (low false positive rate)
  - However, worst at identifying the people that do actually default (high false negative rate)
- **Decision tree has the best recall**
  - Does the best at identifying the people that do actually default
- Does the bank prefer precision or recall?

Topic 1:

What do these algorithms tell us about which features are most important to predict defaulting?

- Use permutation feature importance on decision tree classifier
- **Conclusion: Whether the applicant’s credit is provided by accrediting agency Equinox versus other accrediting agencies is most important feature**
  - Indicates Equinox credit reports might be more accurate than other agencies
- **Other factors include:**
  - Interest rate spread
  - Where application was submitted
  - Charges up front
  - Rate of interest



Topic 2:

- Do social factors play a part in higher interest rates?
- Do these higher interest rates lead to higher default rates?
- Does this suggest some level of discrimination?

Topic 2:

Predict interest rates based on social factors (age, gender, region)

- **Goal: Use KNN Regression to determine if social factors play a role on interest rates**

- **Data Preparation of Social Factors**
  1. **One-hot encoding columns:**
     - **Binned Ages –** Less than 25, 25-34, 35-44, 45-54, 55-64, 65-74, Older than 74
     - **Gender –** Female, Male, Joint (i.e. applied as couple), Gender not specified
     - **Region of US -** North, Northeast, Central, South

- **Result: Predicting interest rate column on social factors leads to *Coefficient of Determination of -0.1***

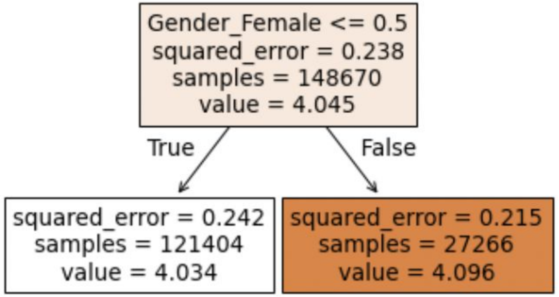
- **Conclusion: No significant correlation between social factors and interest rate**

Topic 2:

Do these higher interest rates lead to higher default rates?

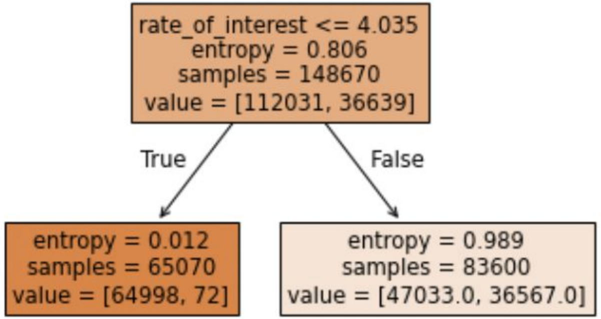
- **Using a Decision Tree Regression, we noticed that female gender was most influential social factor on interest rates**
  - **Females were given higher interest rates than other gender categories**

- **Moreover, a Decision Tree Classifier on rate of interest as feature column and defaulting as target variable indicated the best threshold was an interest rate of 4.035**
  - **interest rates <= 4.035: default rate of 0.1 %**
  - **interest rates > 4.035: default rate of 43.7%**



Other genders' interest rate

Female interest rate

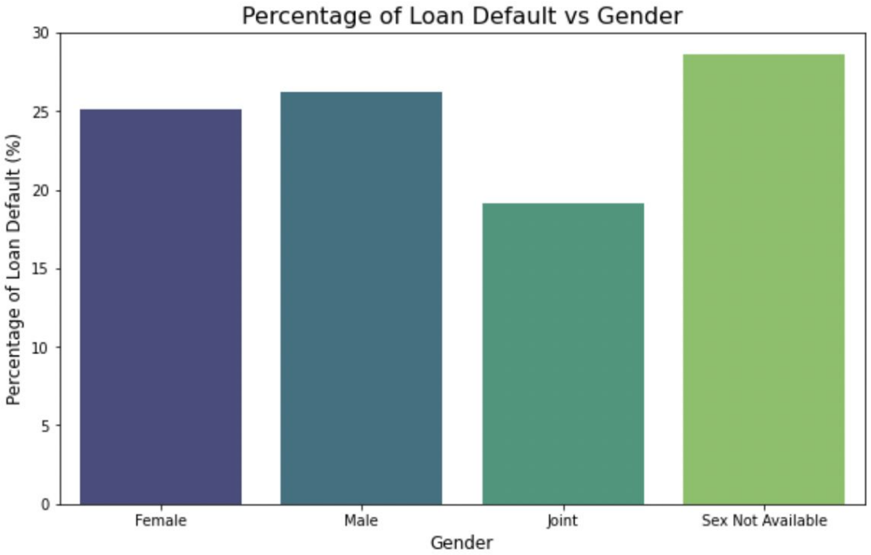


Topic 2:

Does this suggest a level of discrimination that occurs when banks give out loans?

- **Even with females having higher interest rates, on average, males had higher loan default rates**

- **This could suggest that males are more likely to default, possibly because they are being given interest rates they are not qualified for**



Topic 3:

- What variables are most important in having high loan payout?

- What does this say about lenders and you the user?

Topic 3:

What variables are most important in having a high loan payout?

Why is this important?

- **Lets you, be more informed and be able to get a higher loan given out to you, if you want**
- **Higher loan amounts are often riskier, which is an important factor as well**

Steps:

- **Explain data**
- **Data cleaning**
- **Perform Lasso variable selection**
- **Analysis**

Explanation and cleaning: Data on loan defaulting and various parameters and information around those loans.

- **Data has 148,760 rows and 34 columns.***
- **Of those 28 were identified as being potentially relevant, either being a personal characteristic or an aspect of the loan**
- **Ex: Gender, loan type, credit score, loan purpose, business or commercial, etc…**
- **Data had One Hot Encoding performed on it**
- **Used KNN imputation(NN = 3, weight = distance)**

*Large amount of loan data was from conforming loans, which have a built in cap

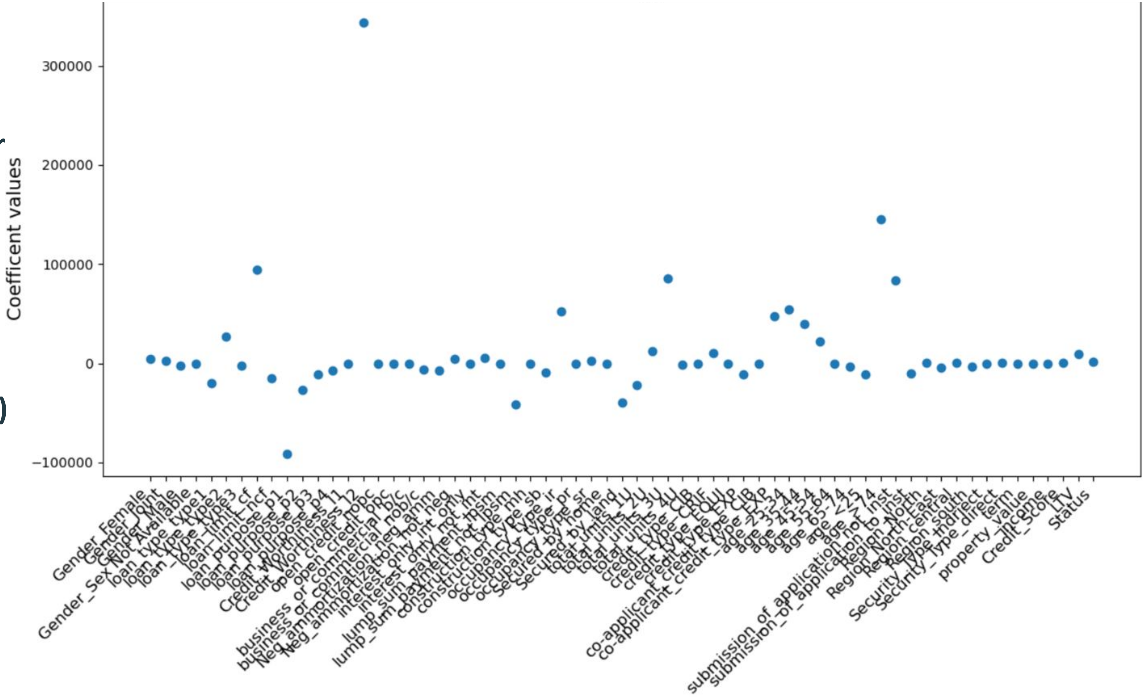
Topic 3:

What variables are most important in having a high loan payout? cont.

Data Modeling:

- Trained Lasso model as a variable selector to discover important factors for identifying high loan payout
- Result: Clear outliers that positively or negatively influenced result
  1. Likeliness to repay a loan
  2. Whether loan is nonconforming
  3. If loan type is highest priority (P1)
  4. Going through a private entity
  5. Being 25-34 years old
  6. Being older than 74
  7. Having 3 units in your property
  8. Seeking funding for investment property

Lasso coefficients mapped to column

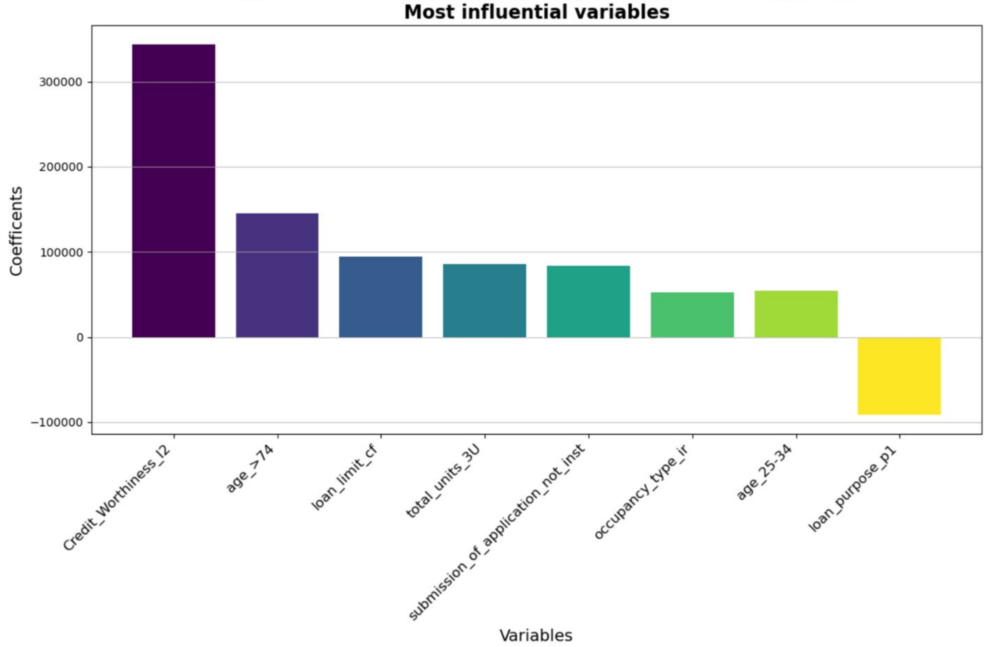


Topic 3:

What variables are most important in having a high loan payout? cont.

What does this indicate about how loan amount is determined?

1. On one hand, might indicate a predatory lending approach that targets those have little experience, are elderly (typically with no source of income), and use private entities who are less regulated
2. On the other, might indicate lenders simply offer good deals to those who aren't in debt and likely to pay their loans back, have built up wealth over life, or are taking out their first loans at an age where they are maturing in their career



Conclusion

Conclusion

01 Algorithm Performance and Recall

- 5-depth decision tree has the best recall
  - Ideal for identifying defaulters
- Logistic regression is precise
  - But struggles with false negatives

02 Key Factors in Loan Default

- The accrediting agency, interest rates, and upfront charges are most influential

03 Insights on Loan Interest and Social Factors

- Females receive higher interest rate but default less than males
  - Hinting at lending gender bias

04 High Loan Payout Variables

- Key Factors: loan type, age, and loan source
  - Potential predatory practices
  - Low risk lending

?

?

Q&A

?

?

THANK YOU!