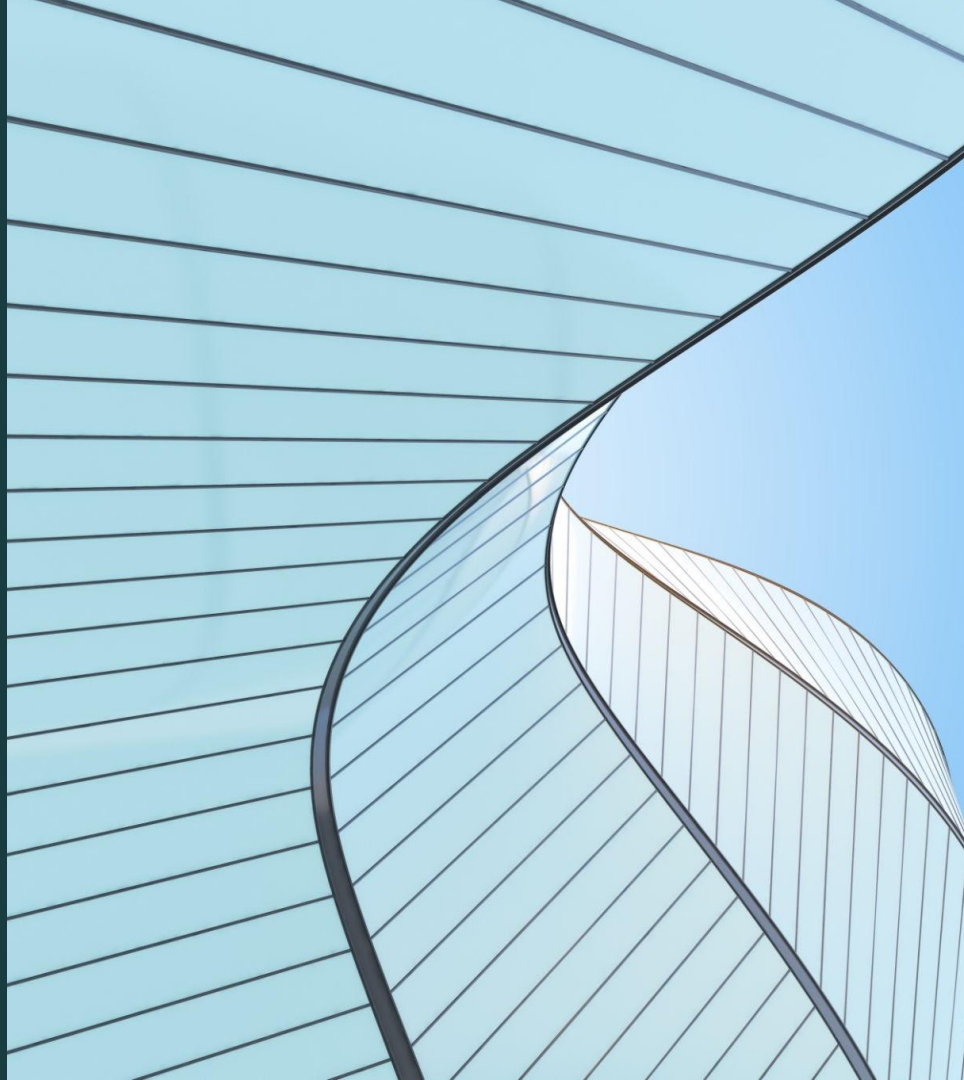
River Magee
5.1.25

Predicting Loan Approval Probability Using Public Lending Data



Business Question

What borrower and loan characteristics are most predictive of loan acceptance?

Build

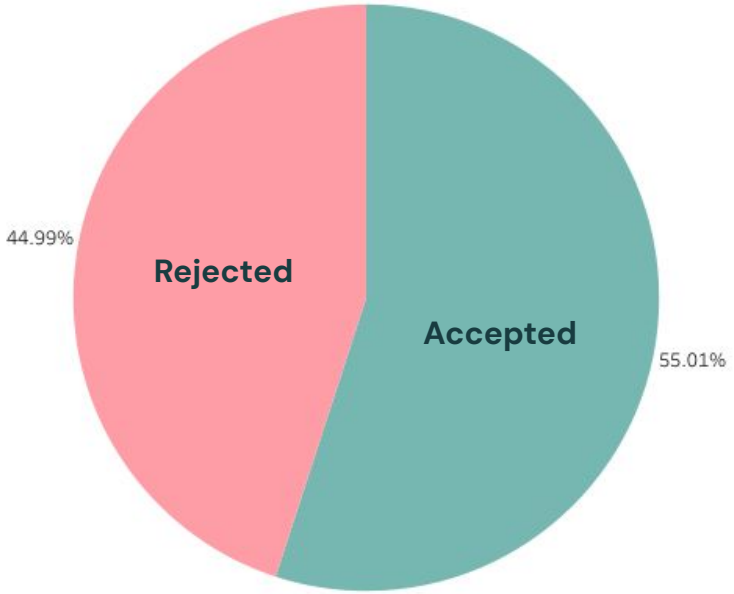
Build a predictive model to estimate the probability of loan acceptance.

Identify

Identify key variables (credit score, income, loan amount, purpose, etc.) that influence loan risk.

Analyze

Segment borrowers by risk to simulate how a financial institution might optimize its approval and pricing strategies.



All Lending Club Loan Data

2007 through 2018 Lending Club accepted and rejected loan data

Total Observation Entries: 2.45 Million

Data Cleaning *overview*

1

Handling missing values

2

Converting categorical variables into factors

3

Standardizing numerical data

4

Removing irrelevant or redundant columns

5

Creating new features (e.g., DTI.)

6

Rule out erroneous data

VARIABLES IN MY ANALYSIS

Between both data sets, 4 common variables existed for comparison.

DTI

Debt To Income Ratio

Risk Score (FICO)

Credit score used as a predictor of default

Loan Amount

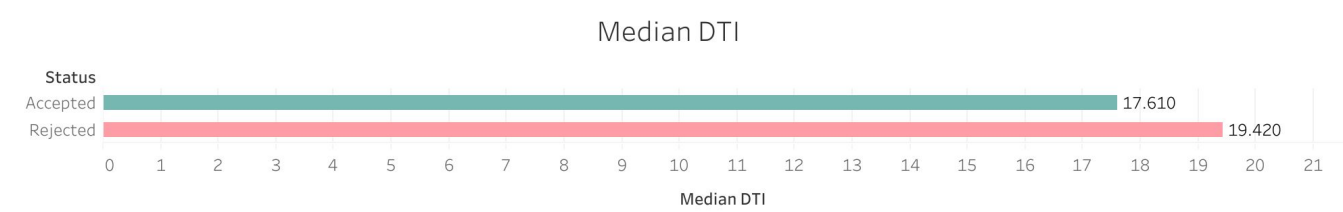
The amount of USD requested in the loan

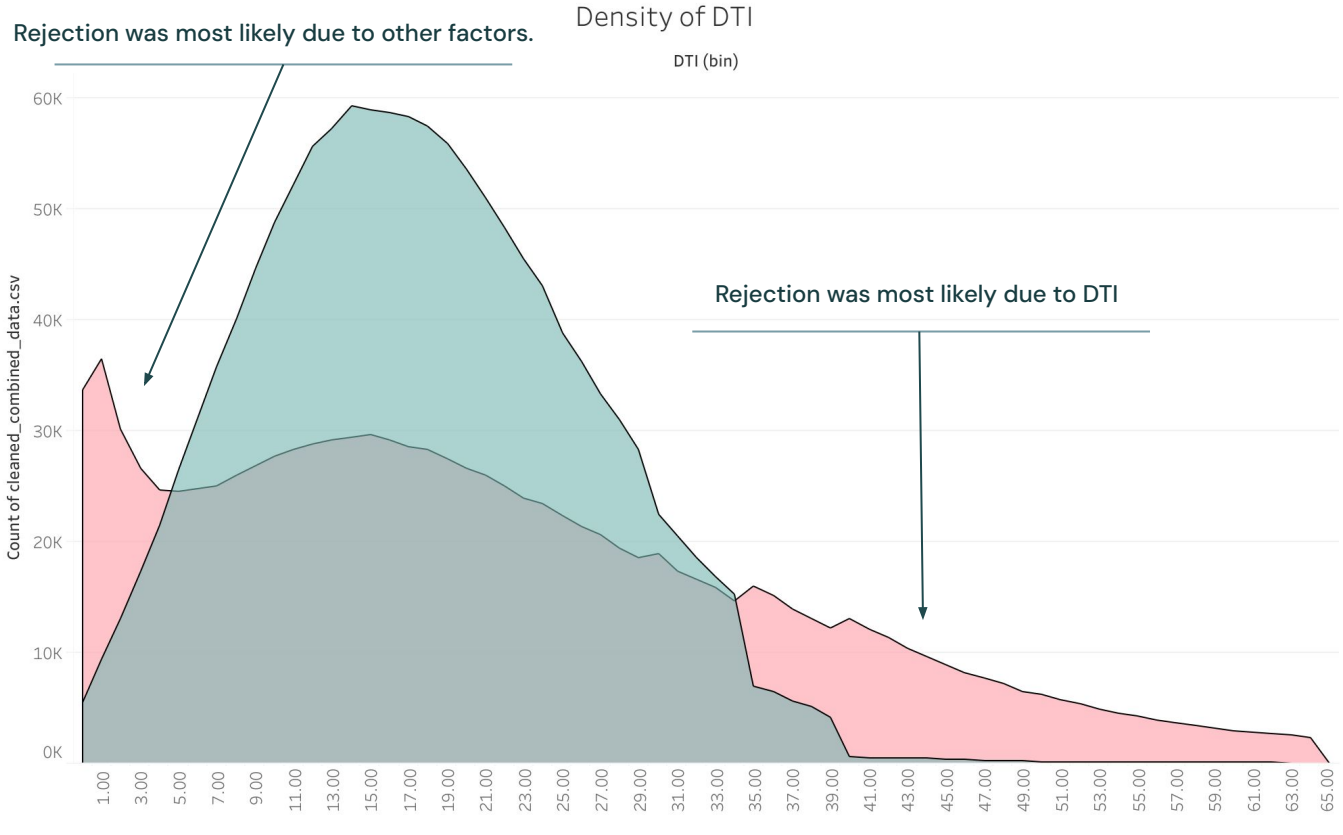
Employment Length

The period of employment before loan application

DEBT TO INCOME RATIO

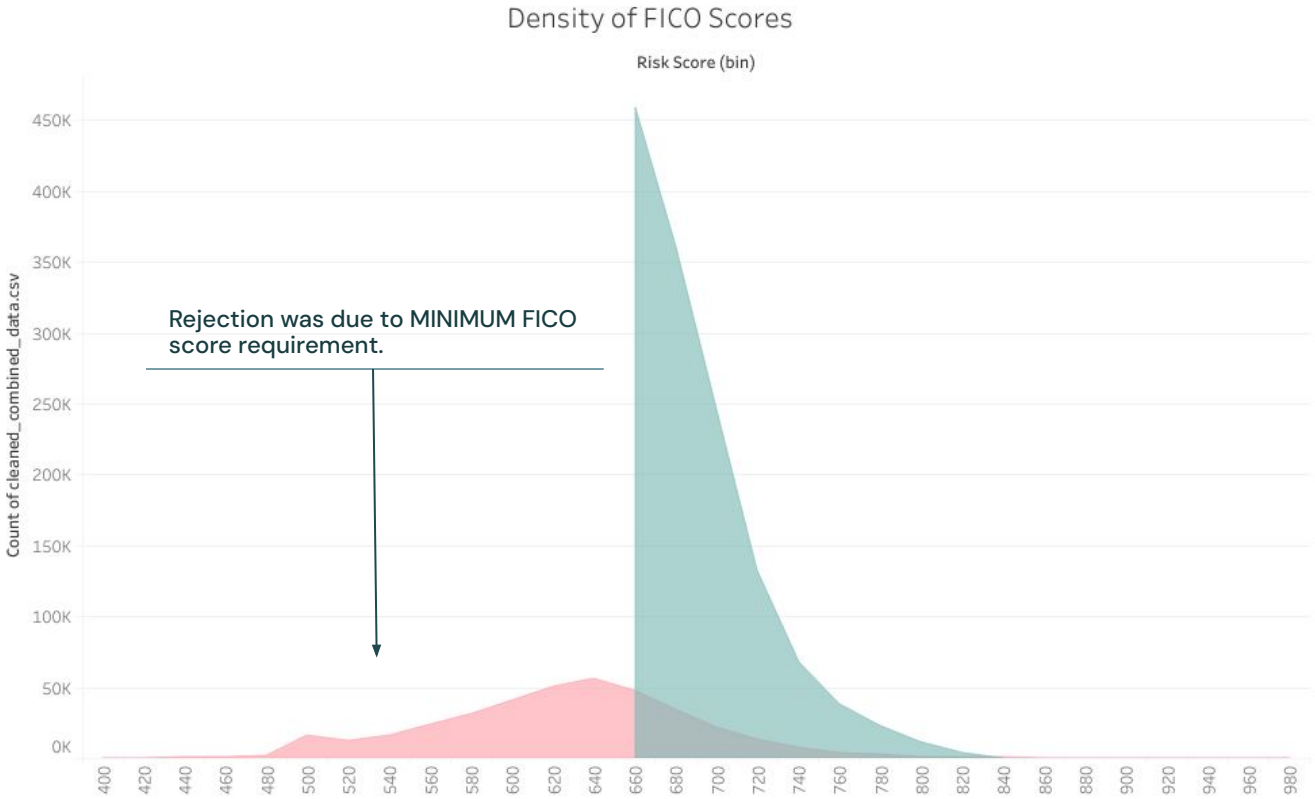
The median of rejected loan applicant DTI Ratios are about **2% higher** than approved applicants.





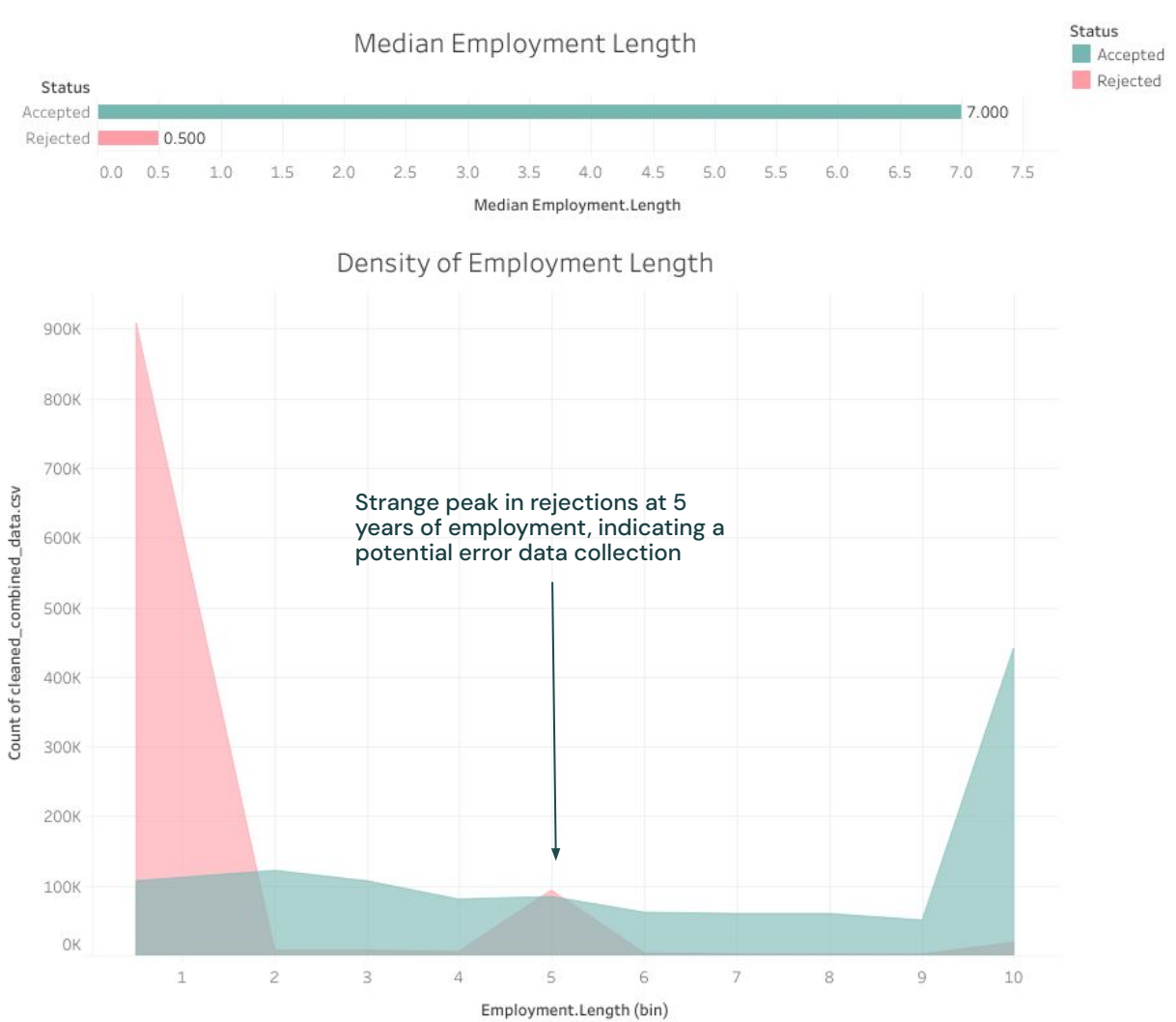
FICO SCORES

The median of rejected loan applicant FICO Scores is below the **MINIMUM** requirement of 660.



EMPLOYMENT LENGTH

The median length of employment for rejected loan applicant is below **one year**, with the vast majority of rejected applicants falling into this range.



Logistic Regression Model Accuracy

Accuracy	**91.7%**	**Overall, the model predicts correctly 91.7% of the time.**
Sensitivity (Recall)	78.9%	78.9% of actual rejections were correctly identified.
Specificity	95.8%	95.8% of actual acceptances were correctly identified.
Precision	85.9%	Of the applications predicted as rejected, 85.9% were truly rejected.
Kappa	0.7688	Substantial agreement between predictions and actual classes.
Balanced Accuracy	87.4%	Average of sensitivity and specificity — good balance.

Confusion Matrix Results

	Predicted: Rejected	Predicted: Accepted
Actual: Rejected	59,661	15,936
Actual: Accepted	9,792	225,570

Major Takeaways

1

The model is especially **strong at correctly identifying accepted** applications.

2

It performs very well on rejections too, though there's a bit more **room for improvement** there.

Understanding the Odds Ratios

Intercept	~9.33e-09	Baseline odds of rejection when all predictors are zero — not meaningful on its own.
Amount.Requested	~0.99998	Slightly decreases odds of rejection with every dollar more requested.
Debt to Income Ratio	~0.9646	Each % increase in DTI slightly reduces odds of acceptance.
Employment.Length	~1.858	Longer employment increases odds of acceptance — very strong predictor.
FICO Score	~1.028	Each point increase in risk score slightly increases odds of acceptance.
AUC Score	**0.9414**	**An excellent score, indicating the logistic regression model has very strong discriminative power.**

Major Takeaways

1

Every year of employment nearly **doubles** the odds of loan approval.

**Could be disrupted by 5 year rejection peak*

2

Every FICO Score point roughly equates to a **3% increase** in approval likelihood.

3

Every 1% increase in DTI **decreases** the odds of loan acceptance by about **3.6%.**

4

The model suggests larger loans are **slightly** more likely to be rejected.

DISCOVERIES

Overall Insights

MOST SIGNIFICANT INDICATOR

Employment length is by far the largest indicator of loan approval using the given data.

01

VISUALIZATIONS MATTER

Using Tableau, we were able to clearly see the required FICO Score for approval, where DTI was likely the reason for rejection, and the distribution of employment length for each category.

02

ROOM FOR IMPROVEMENT

While the model was 95% accurate in predicting approved loans, only 78% of rejected loans were accurately predicted. By continuing to refine our model and the variables it measures, we can aim to boost this accuracy.

03

POTENTIAL APPLICATION

A further refined version of this model could be used to pre-screen loan applicants, informing them of the likelihood of their approval based on their provided factors. This would improve efficiency for financial institutions, perhaps allowing an algorithm to automate minor loan requests, or at the very least flag applicants that are extremely likely to be approved or rejected..

04

FUTURE POTENTIAL

Looking Ahead

Logistic Regression was one way to approach a predictive model for loan approval rates; however, other models like **Random Forests** may provide even higher predictive potential due to their ability to handle complex data, potential variable interactions, and nonlinear relationships.

If this were a full study, testing other predictive models would be essential to ensure the most accurate insights.

Use loan default data in conjunction

By leveraging both loan default and approval data, business could automate reports based on an applicants likelihood of being approved, and if so, a personalized interest rate based on likelihood of default.

Understanding exterior factors

This study was quantitative in nature, but loan risk factors require qualitative assessments as well. An applicant may appear approvable on paper, but there may be other unaccounted variables that increase the chances of default.

Ensuring relative accuracy

While shooting at 92% in the NBA would cement you as the greatest player of all time , an error margin of 8% could be devastating when involving lending approval. Further work may need to be completed before a model like this can be used for practical application.

Thank you.

GitHub R Code: R Programming
Tableau Visualizations: Dashboard
Data Source: Lending Club Data

River Magee