Predicting Credit Default

Kojo Manu
UID: 206545334
Econ 425T
Winter 2025

Problem Identification

Economics/Research Question
Can we accurately predict if someone will default on a loan based on their credit history?

Background/Context
Financial institutions take default risk into account when issuing loans and determining interest rates. Loan defaults can result in significant financial losses, making risk minimization a key priority for lenders. Accurately predicting whether a borrower will default enables financial institutions to make more informed lending decisions, improve risk management, and reduce potential losses. Machine learning models offer a useful approach to use borrowers credit history and enhance traditional credit risk assessment methods.

Data Overview

The data used in this project is the Credit Risk Dataset on Kaggle, which is a dataset containing variables that simulate credit bureau data. This dataset is ideal to answer my research question because it has variables on borrower demographics like age and income, variables on borrower credit history like previous defaults and length of credit history, as well as variables about the loan like the interest rate and the purpose of the loan. All variables included in the dataset are:

Variable	Description	Variable	Description
person_age	Age	loan_amnt	Loan Amount
person_income	Annual Income	loan_int_rate	Loan Interest Rate
person_home_ownership	Home Ownership	loan_status	Loan Status (0=non default, 1=default)
person_emp_length	Employment Length (Years)	loan_percent_income	% Income
loan_intent	Loan Intent	cb_person_default_on_file	Historical Default
loan_grade	Loan Grade	cb_person_cred_hist_length	Length of Credit History

Methodology

1. Logistic Regression Model

I chose to use classification models because the dependent variable loan_status is binary: *loan_status=0* means a non-default loan, and *loan_status=1* means a default loan. I used logistic regression as a baseline model because it is simple, interpretable, and serves as a benchmark to assess whether more complex machine learning techniques can improve prediction performance.

2. Basic Decision Tree Model

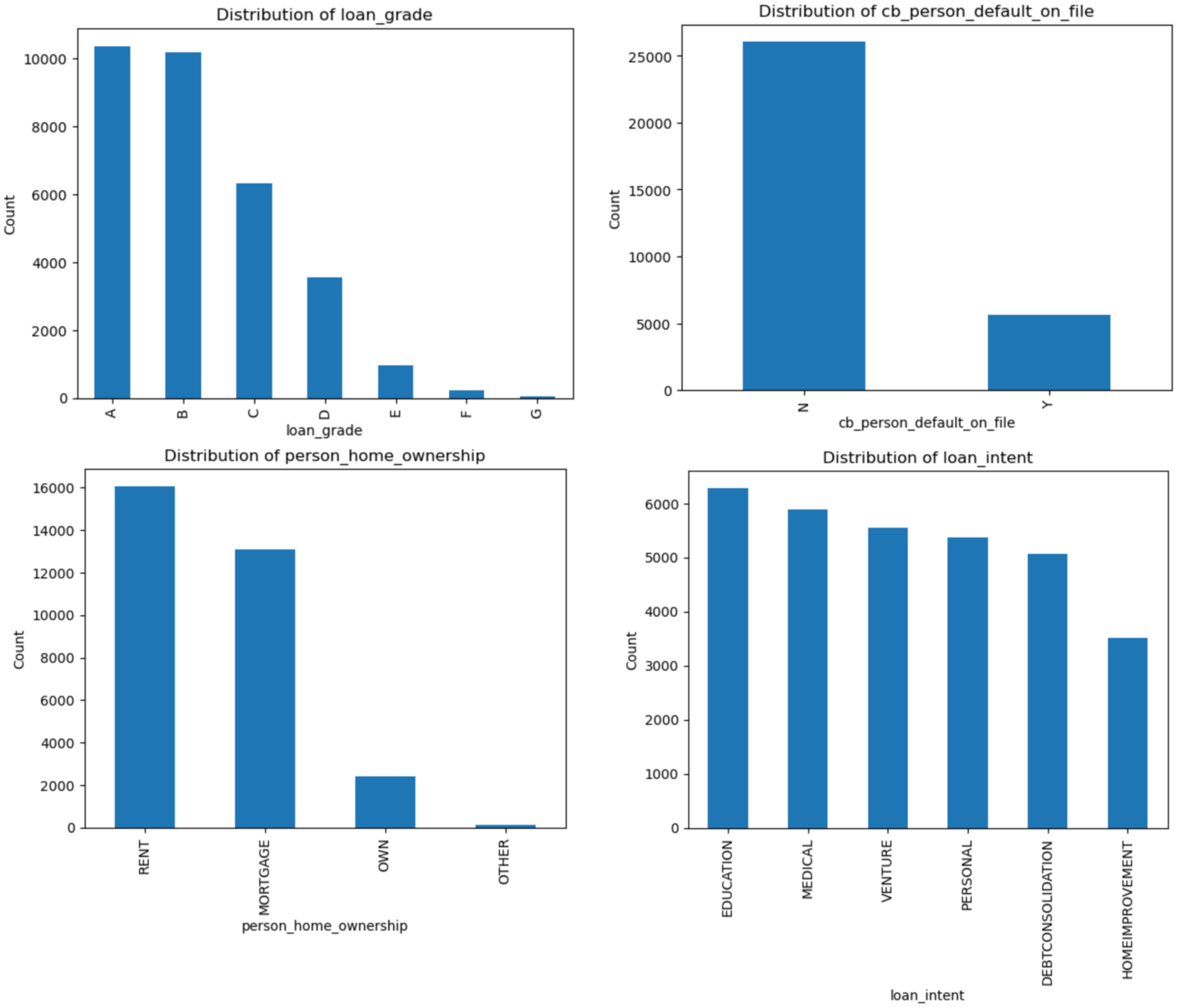
The basic decision tree model was the first alternative I used to compare against logistic regression. I chose a decision tree because it can capture non-linear relationships and interactions between variables, which logistic regression may miss.

3. Cost Complexity Pruned Decision Tree Model

The final model I compare is the cost-complexity pruned decision tree. Pruning helps prevent overfitting by removing unnecessary branches while maintaining high accuracy, making it an improvement over the basic decision tree model.

Data Processing

I first examined the descriptive statistics of the data and checked for any missing values. I then removed extreme outliers and inspected the categorical variables. After analyzing their distributions, I applied one-hot encoding to convert them into numeric format, ensuring compatibility with the models.



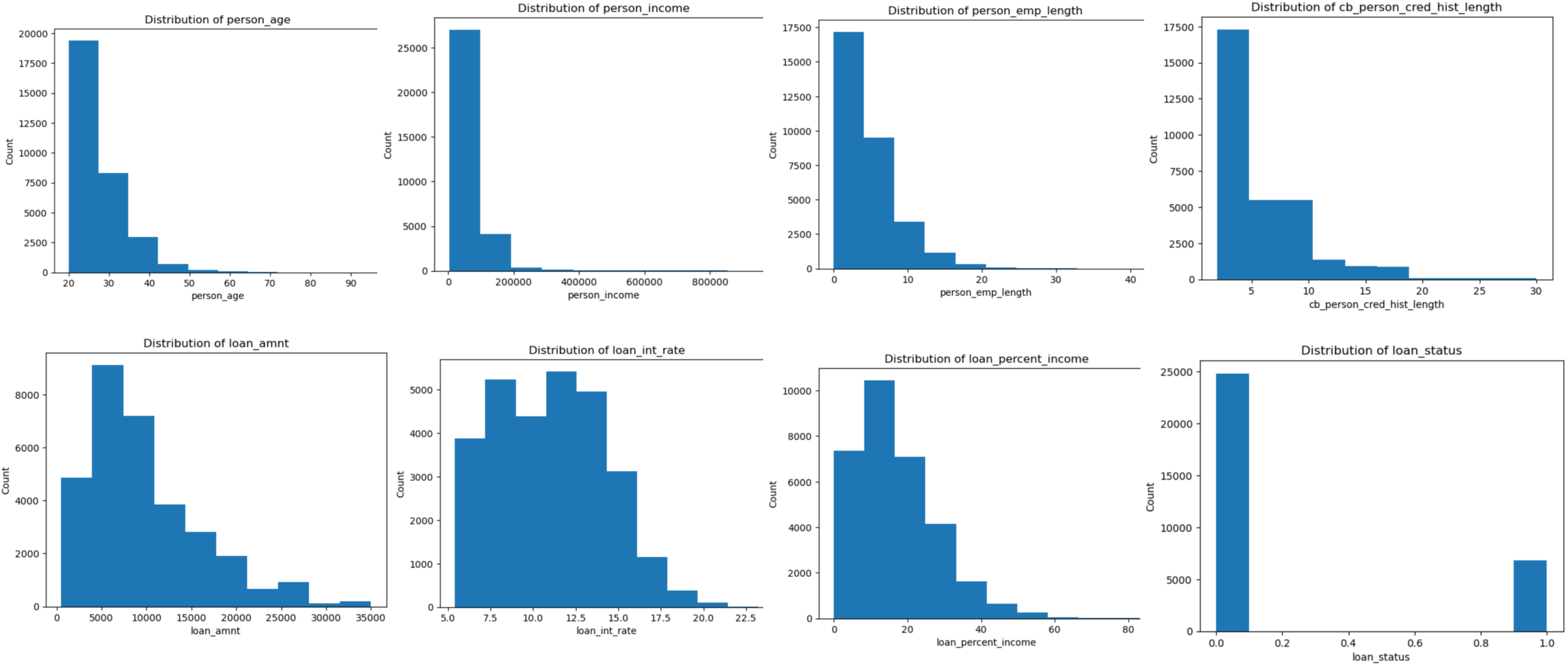
Data Processing

I then looked at the distribution of the numeric columns which can be seen on the next slide. The first thing I noticed was that all the continuous variables were heavily right skewed. Because of this I considered performing log/square root transformations, however these transformations would only benefit the logit model as the decision tree models are not sensitive to the scale of the parameters, so I decided to leave their original distributions.

The second thing I noticed was that the dependent variable, *loan_status*, only had a default rate of 21.55, meaning the dataset is imbalanced with more instances of non-defaults than defaults. This could have potential negative effects on the models like predicting no default too often. I kept this in mind when building the models.

Lastly, I checked for missing values and found that *loan_int_rate* was the only variable with missing entries. To address this, I imputed the median interest rate.

Distribution of Numerical Variables



Model Building

Logistic Regression

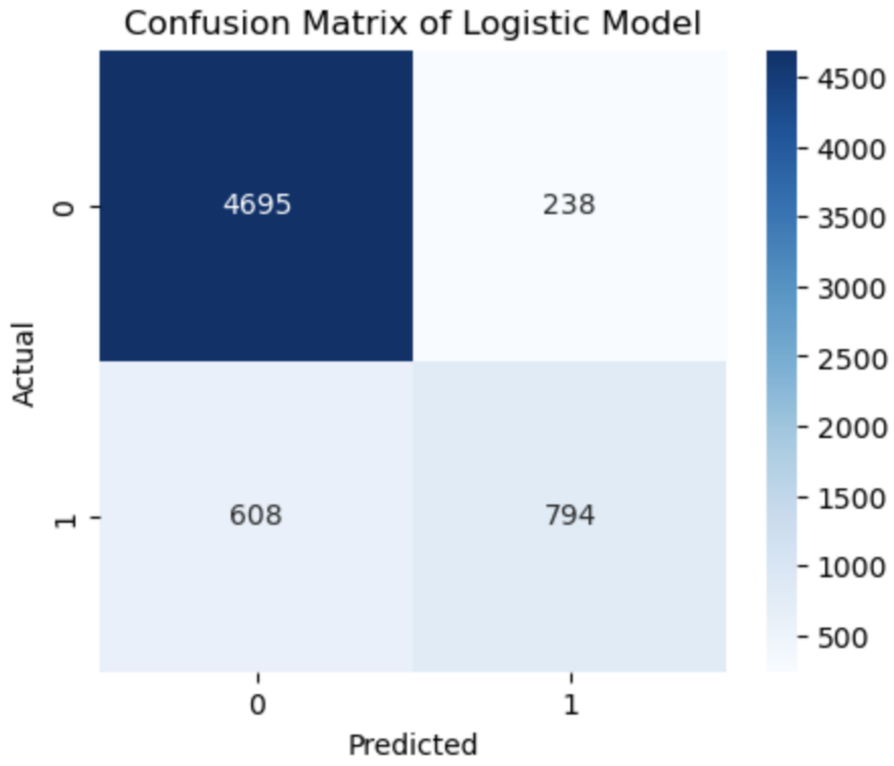
This logit model seems to do an ok job at predicting credit default, with an accuracy of 86.65%. It performs significantly better at predicting "safe" loans than "risky" loans. When predicting a safe loan, it is correct 89% of the time and correctly predicts 95% of safe loans. However, when predicting loans that will default, it is correct 77% percent of the time and only predicts 57% of actual loan defaults. We will see if the basic decision tree can improve on these results.

```
Final Accuracy: 0.8665
Error Rate: 0.1335

Classification Report Logistic:
              precision    recall  f1-score   support

           0       0.89      0.95      0.92      4933
           1       0.77      0.57      0.65      1402

    accuracy                           0.87      6335
   macro avg       0.83      0.76      0.78      6335
weighted avg       0.86      0.87      0.86      6335
```



Summary Results of Logistic Regression

Almost all variables are statistically significant at the 5% level when predicting if a loan will default. The five that are not significant are *person_age*, *cb_person_cred_hist_length*, *loan_intent_HOMEIMPROVEMENT*, *person_home_ownership_OTHER*, and *cb_person_default_on_file_Y*. The one I find most surprising is *cb_person_default_on_file_Y*, because this variables shows if someone has previously defaulted on a loan. I would have predicted that someone who has previously defaulted on a loan would be more likely to default on other loans.

Logit Regression Results

Dep. Variable:	loan_status	No. Observations:	25336
Model:	Logit	Df Residuals:	25313
Method:	MLE	Df Model:	22
Date:	Fri, 21 Mar 2025	Pseudo R-squ.:	0.3571
Time:	01:39:22	Log-Likelihood:	-8457.9
converged:	True	LL-Null:	-13156.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.3622	0.080	-29.558	0.000	-2.519	-2.206
person_age	-0.0136	0.042	-0.325	0.745	-0.095	0.068
person_income	0.1487	0.034	4.393	0.000	0.082	0.215
person_emp_length	-0.0770	0.022	-3.497	0.000	-0.120	-0.034
loan_amnt	-0.7213	0.035	-20.566	0.000	-0.790	-0.653
loan_int_rate	0.1940	0.044	4.432	0.000	0.108	0.280
loan_percent_income	1.4837	0.034	43.274	0.000	1.417	1.551
cb_person_cred_hist_length	-0.0080	0.041	-0.194	0.846	-0.089	0.073
loan_intent_EDUCATION	-0.8222	0.065	-12.690	0.000	-0.949	-0.695
loan_intent_HOMEIMPROVEMENT	0.0466	0.072	0.643	0.520	-0.095	0.188
loan_intent_MEDICAL	-0.1427	0.062	-2.317	0.021	-0.263	-0.022
loan_intent_PERSONAL	-0.6621	0.067	-9.920	0.000	-0.793	-0.531
loan_intent_VENTURE	-1.1127	0.071	-15.723	0.000	-1.251	-0.974
person_home_ownership_OTHER	0.5154	0.322	1.601	0.109	-0.116	1.146
person_home_ownership_OWN	-1.6832	0.118	-14.266	0.000	-1.914	-1.452
person_home_ownership_RENT	0.8539	0.046	18.587	0.000	0.764	0.944
cb_person_default_on_file_Y	0.0054	0.057	0.095	0.924	-0.106	0.116
loan_grade_B	0.2511	0.072	3.494	0.000	0.110	0.392
loan_grade_C	0.4916	0.102	4.842	0.000	0.293	0.691
loan_grade_D	2.5824	0.124	20.909	0.000	2.340	2.824
loan_grade_E	2.7301	0.161	17.010	0.000	2.416	3.045
loan_grade_F	3.3147	0.244	13.603	0.000	2.837	3.792
loan_grade_G	6.7445	1.041	6.478	0.000	4.704	8.785

Model Building

Basic Decision Tree

The basic decision tree improves on the logistic model in certain categories, with a better overall performance. First, it has a higher accuracy of 88.79%. When predicting safe loans, this model is correct 93% of the time and predicted 92% of all safe loans. It also has a corresponding f1-score of 0.93. Like the logistic model, the basic decision tree model has a lower precision when predicting a loan will default than when predicting a loan is safe. When predicting a loan will default, it is correct 74% of the time. There is significant improvement in recall when predicting a loan will default, going from 57% in the previous model to 76% in this model. The f1-score when predicting a loan will default is 0.75. The node count for this decision tree is 4,553.

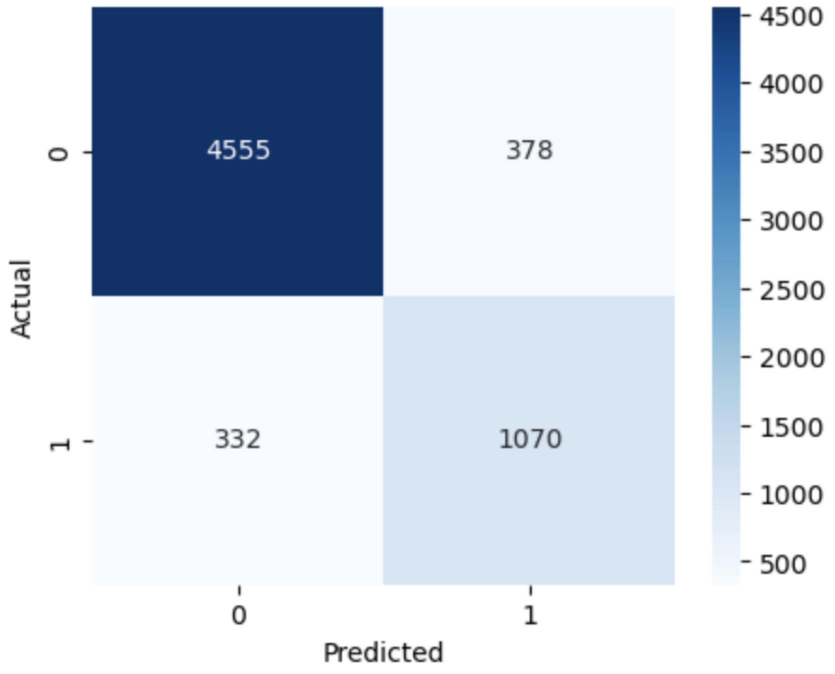
```
Final Accuracy: 0.8879
Error Rate: 0.1121

Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.92      0.93      4933
           1       0.74      0.76      0.75      1402

    accuracy                           0.89      6335
   macro avg       0.84      0.84      0.84      6335
weighted avg       0.89      0.89      0.89      6335
```

Confusion Matrix of Basic Decision Tree Model



Model Building

Cost Complexity Pruned Decision Tree

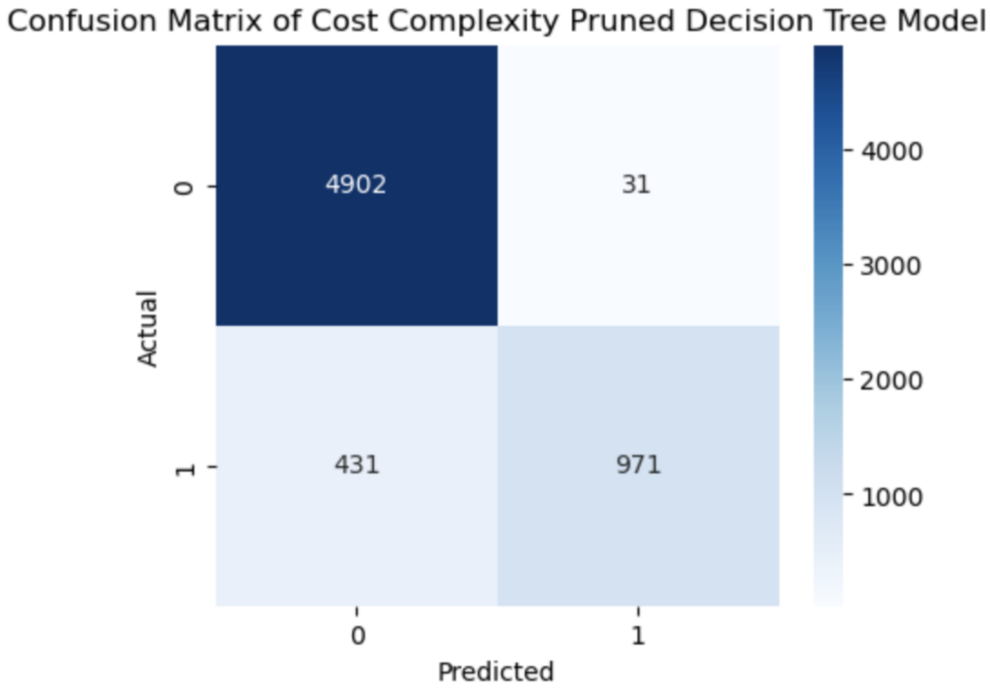
This decision tree model with cost complexity pruning further improves on the basic decision tree model. The optimal alpha is determined to be 0.00025, and it has an accuracy of 92.71%. When predicting a loan is safe, it is correct 92% of the time and predicts 99% of all safe loans with an f1-score of 0.95. When predicting a loan will default, this model is correct 97% of the time but only predicts 69% of all the defaulted loans which is less than the basic decision tree. The corresponding f1-score is 0.81. The node count of this decision tree has been pruned to 137.

```
Optimal alpha: 0.00025
Final Accuracy: 0.9271
Error Rate: 0.0729

Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.99      0.95      4933
           1       0.97      0.69      0.81      1402

    accuracy                           0.93      6335
   macro avg       0.94      0.84      0.88      6335
weighted avg       0.93      0.93      0.92      6335
```



Conclusion

Findings/Insights

Based on the results from the previous slides, both decision tree models outperform the logistic regression model. However, the choice of the best decision tree model depends on the lender's objective.

If a lender is risk averse, they would rather misclassify a safe loan as risky (missing out on potential profits) than misclassify a risky loan as safe (which could lead to financial losses). To prioritize risk minimization, lenders would select the model with the highest recall for predicting loan defaults. Recall measures the percentage of actual defaults that the model correctly identifies.

Among the models, the basic decision tree model achieves the highest recall when predicting loan defaults. Although the pruned decision tree model has higher f1-scores and overall accuracy, the basic decision tree sacrifices precision to improve recall. Given a lender's risk-averse stance, they may prefer the basic decision tree model despite its lower precision, as it is better at identifying loans that are at risk of default.

In the opposite case, a lender less sensitive to risk may employ the pruned decision tree model, which could result in higher profits at the cost of higher risk.

Conclusion

Limitations

One of the main limitations of this project is the low default rate in the dependent variable, *loan_status*. Since only 21.55% of loans defaulted, all three models struggled to accurately predict defaults, leading to lower f1-scores for that class. This imbalance likely affected the models' ability to learn patterns for default prediction.

Another limitation is that the dataset is simulated to mimic credit bureau data. As a result, it may not fully reflect real-world lending decisions, borrower behavior, or economic factors that influence loan defaults. This could impact the model’s generalizability and effectiveness in real-world applications.

Conclusion

Improvements

One improvement would be to use real-world loan data from financial institutions or open-source datasets instead of simulated data. This would make the findings more generalizable and reliable in real lending scenarios.

Another possible improvement is adjusting the decision threshold for predicting loan defaults. Lowering the cutoff from 0.5 would increase the number of predicted defaults, which could be especially beneficial for risk-averse lenders aiming to minimize losses.

Finally, implementing more advanced models, such as Random Forest, XGBoost, or deep learning techniques, could enhance prediction accuracy and capture more complex patterns in the data.

Bibliography

Tse, L. (2020). *Credit Risk Dataset* (Version 1). https://www.kaggle.com/datasets/laotse/credit-risk-dataset/data?select=credit_risk_dataset.csv