PROG25211 AI and Machine Learning

Assignment 1: Logistic Regression

Student name: Sophie Wang

Class number: 1251_93623

Date: February 7, 2025

Table of Contents

Assignment 1: Logistic Regression

Introduction

The following logistic regression is performed on the Loan Approval Classification dataset[1] from Kaggle. The question the analysis seeks to answer is whether a person with the following characteristics would be approved for a loan:

- age: 31
- gender: female
- education: master's
- income: 150,000
- employment experience: 3 years
- home ownership: mortgage
- loan amount: 100,000
- loan intent: home improvement
- loan interest rate: 6%
- loan/income ratio: 66.67%
- credit history length: 10 years
- credit score: 720
- previous loan defaults: 0

[1] https://www.kaggle.com/datasets/taweilo/loan-approval-classification-data

Collecting the Data

The dataset showed 13 different variables: age, gender, education, income, employment experience, home ownership, loan amount, loan intent, loan interest rate, loan/income ratio, credit history length, credit score, previous loan default, and loan status (see Figure 1 below). The loan status (whether the person is approved or not for the loan) is the dependent x variable, while the other variables are the independent y variables. No columns were dropped, as all of the independent variables were hypothesized to be contributing factors to loan status. The dataset did not contain any null values; hence, no rows were dropped.

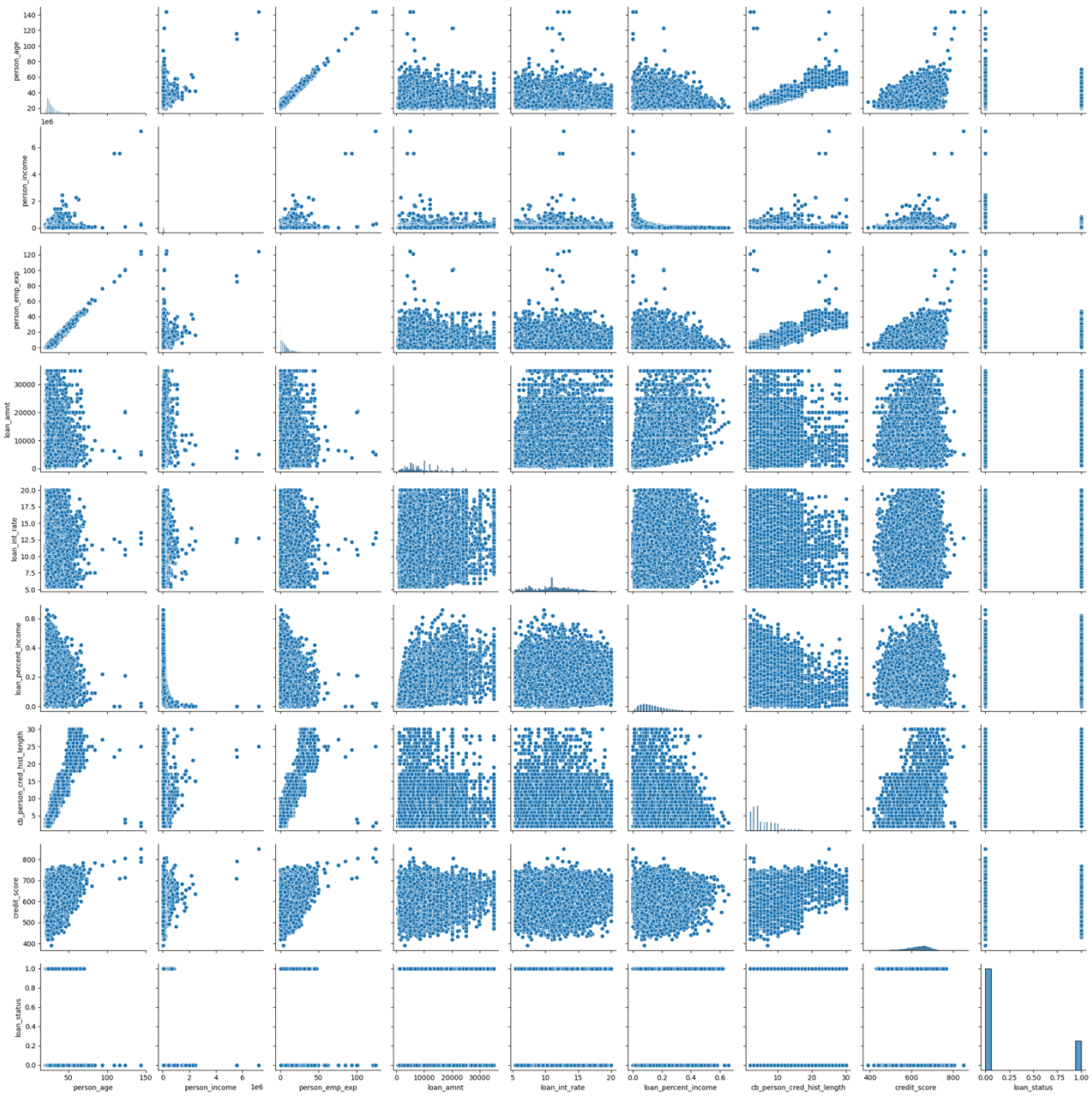
Figure 1: Dataset information

#	Column	Non-Null Count	Dtype
0	person_age	45,000	float64
1	person_gender	45,000	object
2	person_education	45,000	object
3	person_income	45,000	float64
4	person_emp_exp	45,000	int64
5	person_home_ownership	45,000	object
6	loan_amnt	45,000	float64
7	loan_intent	45,000	object
8	loan_int_rate	45,000	float64
9	loan_percent_income	45,000	float64
10	cb_person_cred_hist_length	45,000	float64
11	credit_score	45,000	int64
12	previous_loan_defaults_on_file	45,000	object
13	loan_status	45,000	int64

Data Visualization

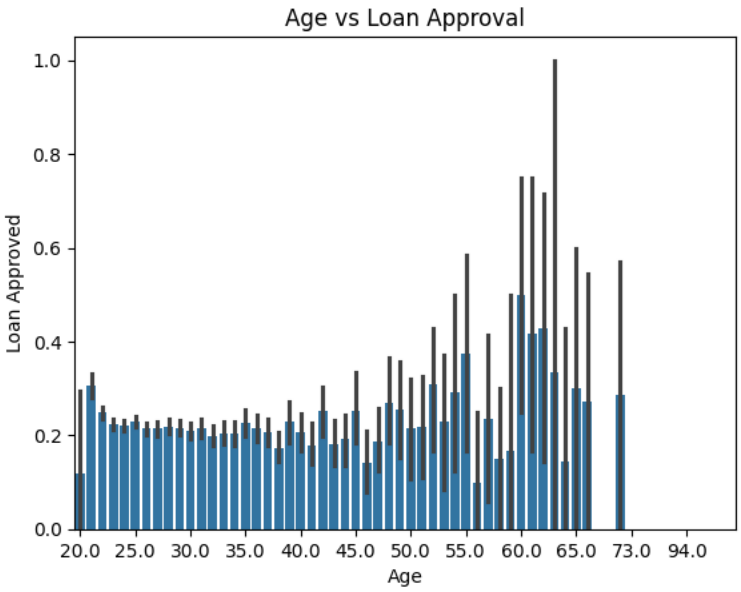
To determine what factors might affect loan status, a pair plot was created (see Figure 2 below).

Figure 2: Pair plot



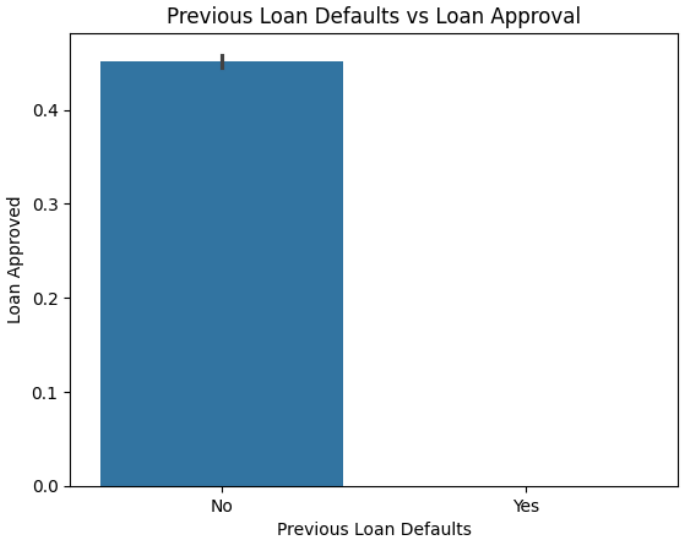
The pair plots show that people who have been approved for loans are in a smaller and lower range of age, income, and employment experience.

Figure 2: Age vs. loan approval



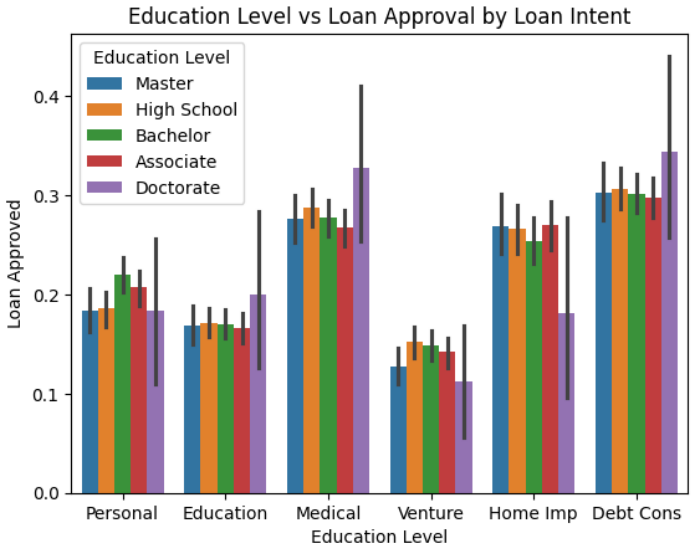
The graph above shows that the likelihood of getting your loan approved is higher at around 21 years old at 30%, drops to about 20% until 35 years old. The likelihood of loan approval also increases in variance with age.

Figure 3: Previous loan defaults vs. loan approval



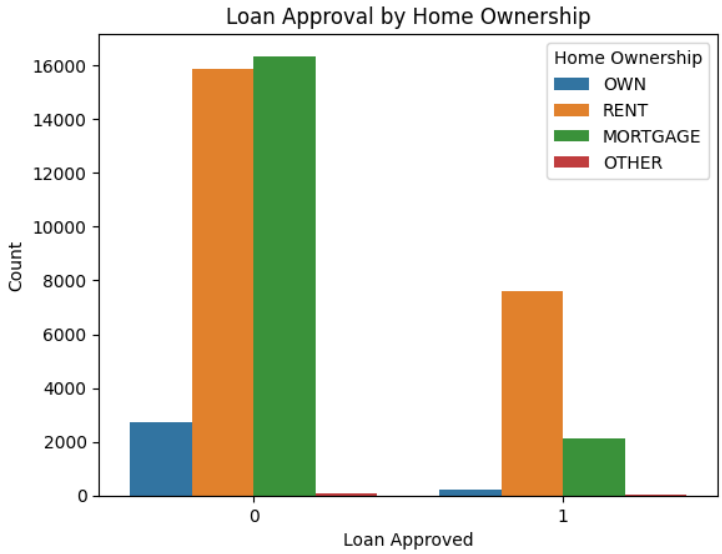
The graph above shows that if there is a previous loan default, the likelihood of loan approval is 0%.

Figure 4: Education level vs. loan approval by loan intent



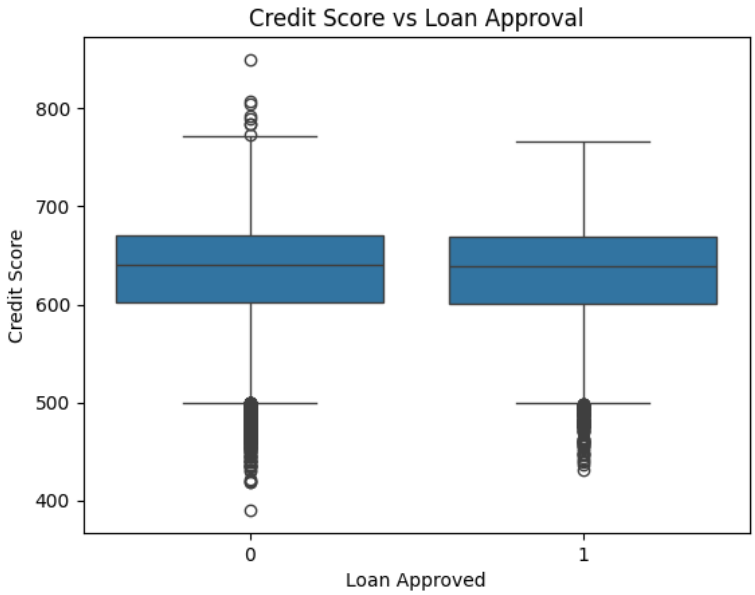
The graph above shows that generally education level is not necessarily associated with a higher chance of loan approval. On the other hand, the loans for ventures are least likely to be approved, while loans for debt consolidation are the most likely, with medical loans being a close second.

Figure 5: Loan approval by home ownership



The graph shows that of the loans that get approved, people who rent are most likely to have their loans approved by far.

Figure 6: Credit score vs. loan approval



The graph shows that the median credit scores are similar for both groups, and the distribution of credit scores is fairly similar for unapproved and approved loans. Surprisingly, there are more extreme outliers for the unapproved group.

Training the Data

Converting Data

```
# convert categorical data to numerical data
data.replace({"person_gender": {"female": 0, "male": 1}}, inplace=True)
data.replace(
    {
        "person_education": {
            "High School": 0,
            "Associate": 1,
            "Bachelor": 2,
            "Master": 3,
            "Doctorate": 4,
        }
    },
    inplace=True,
)
data.replace(
    {"person_home_ownership": {"RENT": 0, "OWN": 1, "MORTGAGE": 2, "OTHER": 3}},
    inplace=True,
```

```
)
data.replace(
    {
        "loan_intent": {
            "DEBTCONSOLIDATION": 0,
            "EDUCATION": 1,
            "MEDICAL": 2,
            "PERSONAL": 3,
            "VENTURE": 4,
            "HOMEIMPROVEMENT": 5,
        }
    },
    inplace=True,
)
data.replace({"previous_loan_defaults_on_file": {"No": 0, "Yes": 1}}, inplace=True)
```

Training the Model

```
# remove dependent data from independent data
X, y = data.drop('loan_status', axis=1), data['loan_status']
# split data into training and testing sets
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,
random_state=42)

from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression

# scale the data
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# train the model
log_model = LogisticRegression()
log_model.fit(X_train_scaled, y_train)
predictions = log_model.predict(X_test_scaled)
```

Evaluating the Model

Classification report results:

Class	Precision	Recall	F1-score	Support
0	0.92	0.93	0.93	10493
1	0.76	0.73	0.74	3007
accuracy	-	-	0.89	13500
macro avg	0.84	0.83	0.83	13500
weighted avg	0.89	0.89	0.89	13500

The classification report shows a high degree of prediction accuracy for predicting when loans are not approved, with 92% precision. On the other hand, the precision is lower, at 76% for approved loans. Overall, the accuracy of the model is 89%. The data we have is imbalanced, since there are more not approved cases than there are approved, which can explain why the model is trained better at predicting unapproved rather than approved loans.

The logistic regression model shows ~89% accuracy score. One suggestion on how to improve the model would be obtain more data for approved loans and retrain the model with more data. Another suggestion could be to use a different type of analysis, such as a linear regression to see if accuracy can be improved. Some columns can also be dropped to test if it can improve the accuracy of the results.

Making a Prediction

The prediction results for the question: This person would likely not be approved for a loan.

Conclusion

The analysis showed approximately 89% accuracy at predicting the loan approval status. It was demonstrated that not having a previous default, renting a home, and borrowing for debt consolidation was the most likely to have the loan approved. Surprisingly, credit scores did not play a huge factor. The model was better at predicting unapproved versus approved loans. Some suggestions for improving the results include obtaining more approved loan data, conducting another type of analysis, or dropping certain variables and retraining.