CISC 482

Computer Systems Seminar

Predicting Credit Card Risk using Machine Learning Models

Author:
Kushal Bhandari

Student number:
843661

Sunday 28th April, 2024

1 Introduction

Broader Context:

In the dynamic world of financial services, accurate credit risk estimation is imperative for banks, lenders, and investors. Amid economic fluctuations, predicting borrower default is essential for sustainable lending. My project focuses on leveraging machine learning and data analytics to enhance credit risk prediction, crucial for informed financial decision-making.

Research Question:

How can machine learning algorithms and data analytics techniques be leveraged to **predict credit risk** effectively, thereby aiding financial institutions in making informed lending decisions while minimizing potential losses?

1.1 Data Description

Description of the how the data was collected

Prof. Hofmann submitted a dataset with 1000 entries that included information on the credit profiles of borrowers. It has 20 categorized/symbolized features, including financial indications like savings, checking account balances, and housing status, as well as personal data like age, sex, and kind of job. It also documents loan details such as credit amount, term, and purpose. Every entry is categorized as either a good or bad credit risk, providing predictive modeling with a well-defined target.

Relevance for Financial Institutions

Financial organizations looking to improve their approaches to credit risk assessment may find great value in this dataset. Institutions can customize lending policies by identifying patterns associated with creditworthiness through the analysis of borrower data. By using machine learning, organizations can reduce risk exposure, improve operational efficiency, and automate credit evaluation—all of which contribute to a healthier lending portfolio.

1.1.1 Meaning of the Variables

- **Age**: Numeric variable representing the age of the individual applying for credit.
- **Sex**: Categorical variable indicating the gender of the individual, with options being "male" or "female".
- **Job**: Numeric variable representing the employment status and skill level of the individual, with categories ranging from 0 (Unskilled and non-resident) to 3 (Highly skilled).
- **Housing**: Categorical variable indicating the housing status of the individual, with options being "own", "rent", or "free".
- **Saving accounts**: Categorical variable describing the amount of savings the individual has, with options including "little", "moderate", "quite rich", and "rich".
- **Checking account**: Numeric variable representing the balance in the individual's checking account, measured in Deutsch Mark (DM).
- **Credit amount**: Numeric variable representing the amount of credit requested by the individual, measured in Deutsch Mark (DM).
- **Duration**: Numeric variable representing the duration of the credit in months.
- **Purpose**: Categorical variable indicating the purpose of the credit, with options including "car", "furniture/equipment", "radio/TV", "domestic appliances", "repairs", "education", "business", and "vacation/others".
- **Risk**: Target variable indicating the credit risk associated with the individual, with options including "Good" for good credit risk and "Bad" for bad credit risk.

1.1.2 Course Relevance

In Business Analytics 310-41, I utilized Microsoft Excel to accurately forecast Springfield College's enrollment trends using linear regression techniques. This honed my data analysis and predictive modeling skills in business and finance. Drawing from this experience, I'm ready to apply my expertise to offer valuable insights and solutions, be it optimizing credit risk assessment, developing predictive models, or implementing innovative risk mitigation strategies.

1.2 Data Description

Age	Sex	Job	Housing	Saving accounts	Checking account	Credit amount	Duration	Purpose	Risk
67	male	2	own	NaN	little	1169	6	radio/TV	good
22	female	2	own	little	moderate	5951	48	radio/TV	bad
49	male	1	own	little	NaN	2096	12	education	good
45	male	2	free	little	little	7882	42	furniture/equipment	good
53	male	2	free	little	little	4870	24	car	bad
35	male	1	free	NaN	NaN	9055	36	education	good
53	male	2	own	quite rich	NaN	2835	24	furniture/equipment	good
35	male	3	rent	little	moderate	6948	36	car	good
61	male	1	own	rich	NaN	3059	12	radio/TV	good
28	male	3	own	little	moderate	5234	30	car	bad

Figure 1: Data Head

2 Exploratory Data Analysis

I'll begin EDA with a statistical overview of the dataset and correlation analysis between predictor variables and 'Risk'. Then, I'll illustrate each predictor's relation to 'Risk' with graphs, including a pairplot for visualization.

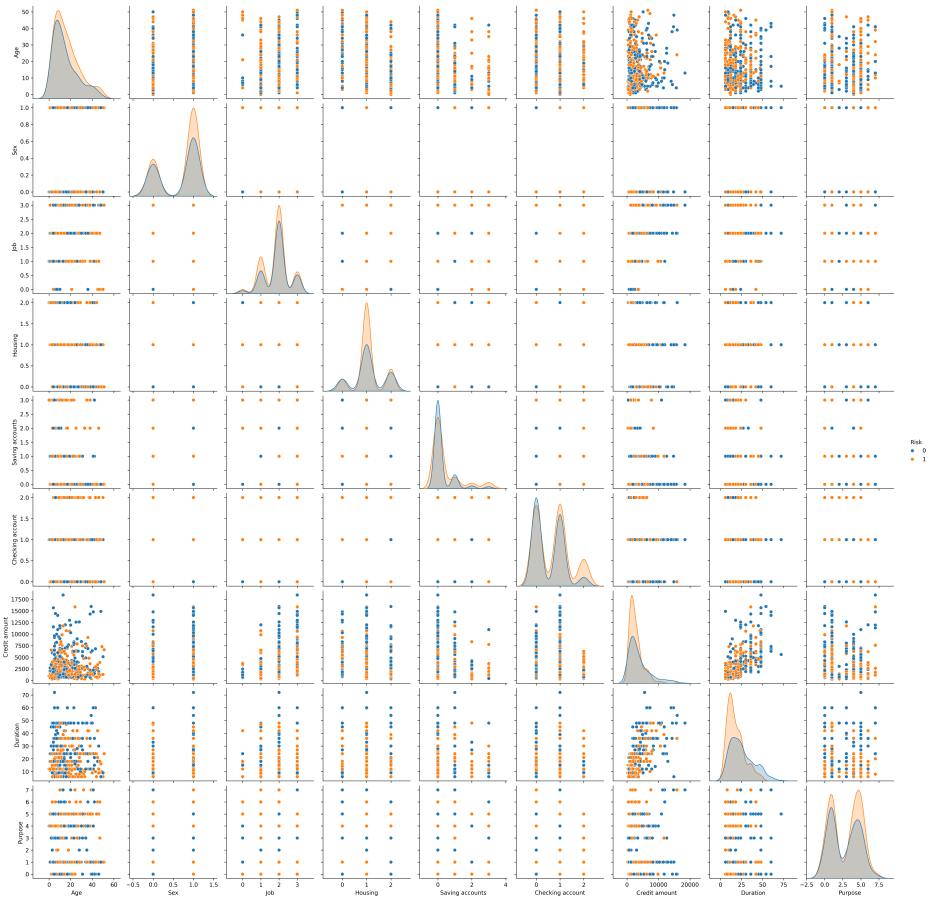

Figure 2: Pair Plot

2.1 Response Variable (Label or Target)

The response variable is the 'Risk' feature. Risk is categorized as "Good," indicating good credit risk, and "Bad," indicating bad credit risk. I am mapping the Risk column values into numerical values of 0 and 1. This is because these numbers represent a natural ordering of the risk levels, where 0 signifies good credit risk, and 1 signifies bad credit risk.

2.2 Predictor Variables (Features)

Correlation Analysis

- **Risk**: 1.000000 (Target variable indicating credit risk)
- **Checking account**: 0.140457
  This feature likely indicates the status of the applicant's checking account, with a positive correlation of 0.140. Individuals with higher checking account balances or better account status may pose lower credit risk.
- **Saving accounts**: 0.127930
  Similar to checking accounts, saving accounts reflect the financial stability of the applicant, showing a positive correlation of 0.128. Higher savings could suggest lower risk due to better financial management.
- **Sex**: 0.063200
  The correlation of 0.063 suggests a weak positive relationship between gender and credit risk. However, interpreting gender-based risk requires caution and ethical consideration, as it may not be directly linked to financial behavior.
- **Age**: 0.055209
  With a correlation of 0.055, age seems to have a slight positive association with risk. Younger individuals may have less established credit histories or financial stability, potentially leading to higher risk.
- **Purpose**: 0.051416
  The correlation of 0.051 indicates a weak positive relationship between the purpose of the loan and credit risk. Certain loan purposes may be riskier than others, depending on factors such as investment returns or repayment likelihood.
- **Housing**: -0.006575
  This feature, with a correlation close to zero (-0.007), shows minimal relationship with credit risk. Housing status alone may not be a strong predictor of credit risk in this dataset.
- **Job**: -0.049555
  The negative correlation of -0.050 suggests a weak inverse relationship between job type and credit risk. More stable or higher-paying jobs may correlate with lower credit risk.
- **Credit amount**: -0.183392
  With a correlation of -0.183, there is a moderate negative relationship between the amount of credit and credit risk. Higher credit amounts may indicate greater financial responsibility or capability to handle larger debts, leading to lower risk.
- **Duration**: -0.293611
  The strongest negative correlation of -0.294 indicates a moderate inverse relationship between loan duration and credit risk. Shorter loan durations may suggest lower risk as they imply quicker repayment and potentially less exposure to financial uncertainties.

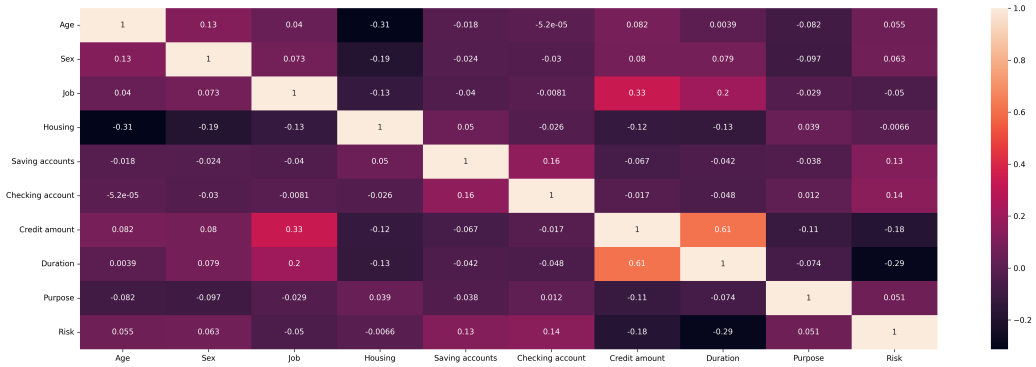

Figure 3: HeatMap

2.3 Histplot

The following figure depicts histplot distributions of predictor variables against the response variable, Risk. The selection of predictor variables—Saving Accounts, Housing, Purpose, Sex, Checking Account, and Job—was informed by the correlation analysis conducted via Heatmap above. These variables were chosen based on their positive correlation with the risk variable.

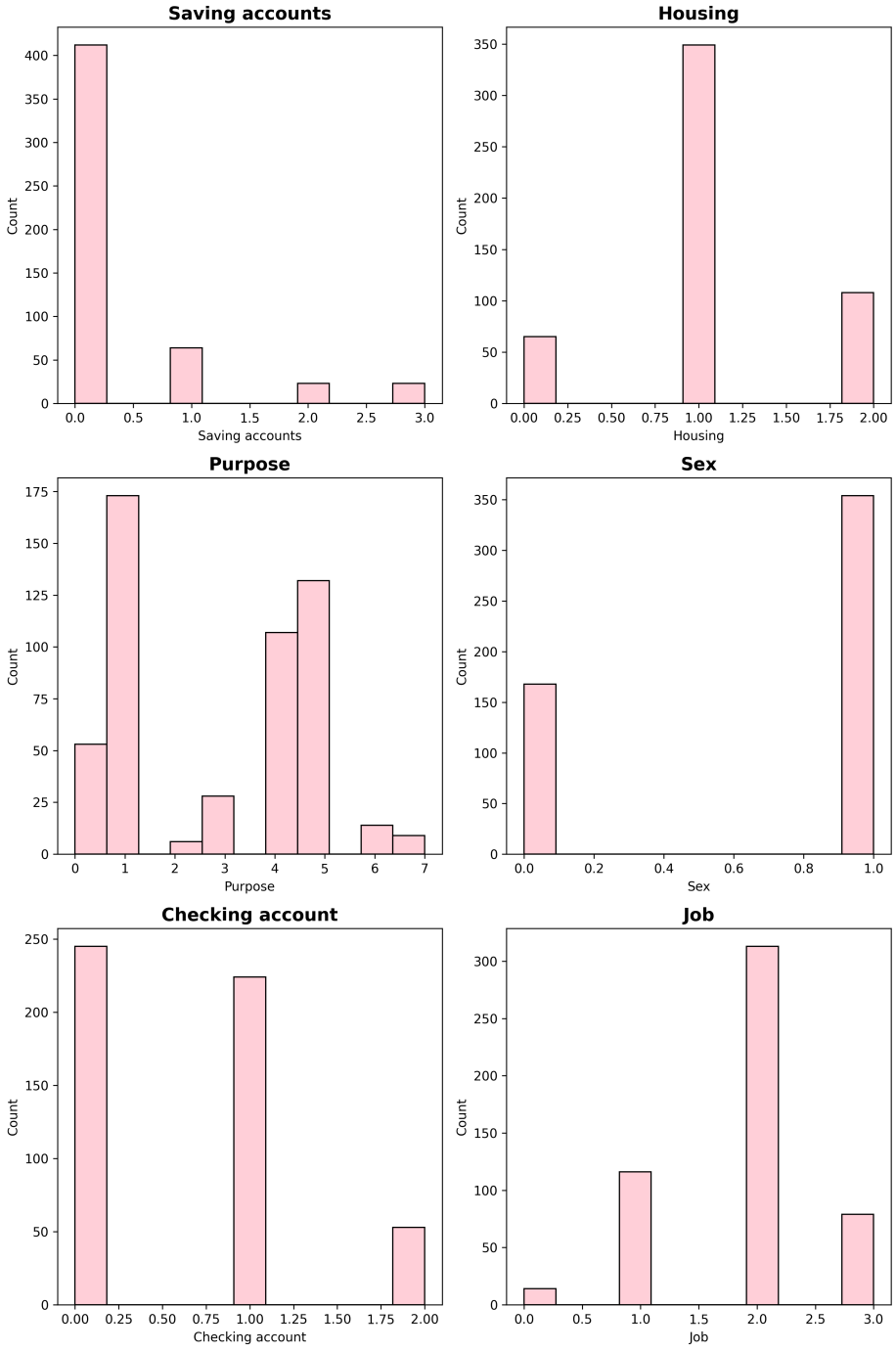

Figure 4: Histplot

2.4 Data Cleaning

In the data cleaning process, null values were addressed within specific columns of the dataset. Zeros were replaced with NaN (Not a Number) in the "Saving accounts" and "Checking account" columns, which

initially had 817 and 606 non-null entries, respectively. Subsequently, rows containing NaN values were removed from the dataset to ensure data integrity for analysis.

2.5 Data Splitting

Since all the predictor features like Saving accounts, Checking account, Housing, Purpose, Job, and Sex are important features with good positive correlation with our response variable 'Risk', I am using these features as our predictor features (X). Then, I am splitting our dataset into train and test dataset to train and finally evaluate our model using the test dataset.

2.6 Data Scaling

Data scaling is done to bring all the features or variables of a dataset to the same scale, so that no one feature dominates the others. Different features can have different ranges of values. So,scaling helps in improving the performance of some machine learning algorithms like KNN and SVM which are sensitive to the scale of the input feature.

3 Data Modeling and Analysis

3.1 Logistic Regression

Logistic regression is defined as a supervised machine learning algorithm that accomplishes Binary Dataset classification. The Decision Tree classifier attained a **training accuracy of 62.19%** and a **testing accuracy of 55.14%** using an **80/20 train-test split**.

3.1.1 Confusion Matrix

The Confusion Matrix is a vital tool in assessing a machine learning model's performance on test data. It categorizes predictions into true positives (correctly identified positive cases), true negatives (correctly identified negative cases), false positives (incorrectly identified positive cases), and false negatives (incorrectly identified negative cases). This breakdown offers precise insights into the model's accuracy and areas for improvement.

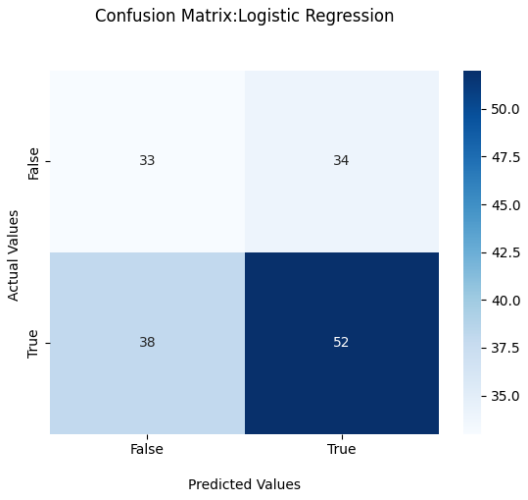

Figure 5: Confusion Matrix

3.2 Decision Tree Classifier

A decision tree classifier is a machine learning model that predicts outcomes by dividing the input space into regions, assigning each region a class label based on features of the input data. The Decision Tree

classifier attained a **training accuracy of 82.41%** and a **testing accuracy of 56.68%** using an **70/30 train-test split**.

	precision	recall	f1-score	support
0	0.49	0.51	0.50	67
1	0.62	0.61	0.62	90
accuracy			0.57	157
macro avg	0.56	0.56	0.56	157
weighted avg	0.57	0.57	0.57	157

3.3 KNN Classification

KNN, short for K-Nearest Neighbors, is a versatile algorithm used for classification and regression tasks. It operates by examining the "k" nearest data points in the training set and making predictions based on either the majority class for classification or the average value for regression. The KNN classifier achieved a **training accuracy of 69.04%** and a **testing accuracy of 60.50%** with a **70/30 train-test split**.

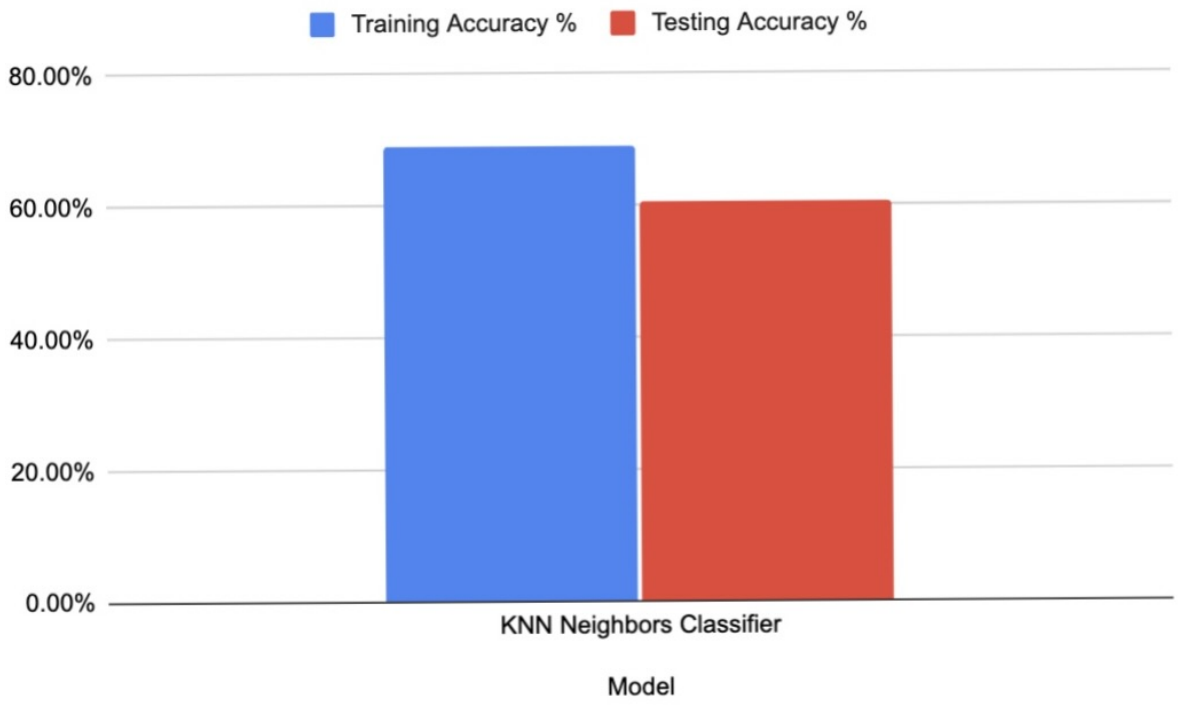

Figure 6: KNN Classifier

4 Future Recommendation

Here are some future recommendations:

- Address outliers in the dataset to enhance model accuracy.
- Optimize hyperparameters and explore alternative algorithms such as Random Forest or SVM to improve predictive performance.
- Improve model interpretability through techniques like feature importance analysis and SHAP values for better understanding and trust.
- Explore individual relationships of predictor variables against 'Risk' and delve into subcategories' relationships with 'Risk' for deeper insights.

5 Discussion

I am pursuing a double major in Business Management and Computer Science, which has provided me with a diverse skill set. Proficient in C++, Python, and MySQL, I am equipped to tackle complex data challenges and derive actionable insights.

In my Business Analytics 310-41 course, I delved into the world of business analysis. Leveraging linear regression techniques, I successfully predicted enrollment trends at Springfield College with precision. This hands-on experience not only sharpened my data analysis skills but also fortified my prowess in predictive modeling, particularly in the realms of business and finance. Working within the International Admissions Office and utilizing Slate Database has streamlined our data management processes, allowing for efficient tracking of applicant information, communication logs, and enrollment statistics. This platform empowers me to navigate through complex datasets seamlessly, enabling me to derive actionable insights and make informed decisions.

This project deeply resonated with my past experiences, aligning perfectly with my passion for utilizing data to inform strategic decisions. Equipped with a computer science degree, I am driven to combine my coding skills with data analysis to revolutionize business practices. My goal is to integrate fintech knowledge into operational processes, whether it's enhancing credit risk assessments, developing predictive algorithms, or spearheading innovative tech solutions to mitigate risks.

Continuous exploration, experimentation, and refinement are key to accurately predicting outcomes. Diverse EDA techniques and machine learning models offer rich opportunities for discovery. I am excited to deepen my understanding and refine skills further.