Data-Driven Credit Risk Modeling: Predicting Probabilities of Default and Assigning Credit Scores

Quynh Lan Nguyen (ID: 501276889)

Supervisor: Tamer Abdou

Date of Submission: January 17th, 2024

Ryerson University

Table of Contents

Abstract

Within the dynamic realm of financial lending, the development of robust credit risk models is vital for promoting responsible lending practices and ensuring financial stability. In my project, I analyze an extensive dataset of over 32,581 consumer loan transactions, featuring 12 distinct attributes. My goal is to utilize Python to craft a data-driven credit risk model that not only predicts the probabilities of default (PD) but also assigns accurate credit scores to borrowers. The model is designed to be transparent and user-friendly, serving as an indispensable tool for financial institutions in making well-informed lending decisions.

Dataset:

This dataset is pivotal in my research, offering deep insights into borrowers' characteristics, payment behaviors, and loan statuses. It's the foundation upon which I aim to build a model that enhances credit assessment precision and encourages responsible lending. Please find the link below to explore the complete dataset:

https://www.kaggle.com/datasets/laotse/credit-risk-dataset

Research questions

1. **Determinants of Default Probability:** What factors are most predictive of loan defaults? I examine borrower data to identify key indicators of credit risk.

2. **Development of an Interpretable Scorecard:** How can we construct a scorecard that transparently assesses credit risk? The approach involves logistic regression, leveraging statistical measures such as information value and weight of evidence.

3. **Model Validation and Reliability:** How dependable is the credit risk model I've developed? I rigorously validate the model's performance using cross-validation, ROC curves, and calibration tests to ascertain its effectiveness in real-world lending scenarios.

Through this project, I aim to deliver a model that not only enhances financial institutions' lending decisions but also serves as a catalyst for risk management, ultimately supporting financial security and ethical lending practices.

Literature review

Title of the Paper: Machine learning-driven credit risk: a systemic review

Authors: Si Shi, Rita Tse1,Wuman, Stefano D'Addona, Giovanni Pau

Summary:

Key Findings: The study under review provides a pivotal examination of current leading technologies in credit risk estimation, particularly focusing on statistical, machine learning, and deep learning techniques. The authors introduce an innovative taxonomy that marries financial concepts with machine learning principles, assessing the efficacy of various methods using public datasets. This approach aligns closely with my project's objective of developing a data-driven credit risk model, especially given the emphasis on deep learning and machine learning techniques.

Methodologies: the paper is grounded in a systematic literature review following PRISMA guidelines, evaluating 76 papers selected through a rigorous filtering process. This methodology

is particularly instructive for my project, offering insights into the systematic evaluation of literature and methodologies.

Insights: The key findings of the paper, highlighting the superiority of deep learning methods and the advantages of ensemble methods in credit risk estimation, are directly relevant to my research. These insights reinforce my decision to explore advanced analytical techniques in my model development. The challenges identified, such as data imbalance and model transparency, are critical considerations for my project.

While the conclusions drawn from the study underscore the potential of machine learning in credit risk modeling, the paper's limitations in scope – specifically its focus on binary classification and lack of discussion on ethical concerns – present areas for further exploration in my project. Additionally, the absence of practical, real-world applications in the paper suggests a gap that my project could aim to fill, particularly in demonstrating the practical applicability of these models.

In summary, this paper offers valuable insights into advanced credit risk estimation methods and presents a well-constructed comparative analysis. Its findings and methodologies will inform several aspects of my project, while its limitations provide a clear direction for further research and practical application.

Title of the Paper: A New Model Averaging Approach in Predicting Credit Risk Default

Authors: Paritosh Navinchandra Jha and Marco Cucculelli

Summary:

<u>Purpose of the Study:</u>

This paper introduces an innovative approach to credit risk modeling, focusing on a new ensemble technique that uses a weighted model averaging strategy. This method, distinct from traditional Bayesian or frequentist techniques, aims to enhance the accuracy of credit risk assessments. This aligns closely with my project's aim of leveraging advanced methodologies to predict credit risk more effectively.

Methodology:

The authors develop an ensemble model through a unique quadratic programming approach, creating a weighted combination of various machine learning models. This method prioritizes minimizing the covariance between individual model errors. Tested on a dataset from a financial institution, it showcases its adaptability, a feature that is particularly relevant to my project given the diverse nature of credit risk data.

Key Findings:

The study finds that this model averaging approach surpasses traditional single-model and other ensemble model performances.

It addresses common challenges in credit risk modeling, such as data imbalance and transparency, which are crucial considerations for my project.

Conclusions:

The paper concludes that combining multiple machine learning models through this averaging technique significantly benefits credit risk modeling. It demonstrates a superior ability to manage complex datasets and capture dynamic relationships, which is vital for my project's goal of developing a robust predictive model.

Analysis:

The insights from this paper are invaluable for my project, particularly in terms of adopting an ensemble approach to improve prediction accuracy. The methodology could potentially enhance my model's performance, considering the complex nature of credit risk data. However, the focus on a specific ensemble method might limit broader applicability, which warrants careful consideration in my project. Additionally, exploring its practical implementation in real-world scenarios would add depth to my project, addressing a gap identified in the paper.

Incorporating this study into my literature review not only underscores the importance of advanced modeling techniques in credit risk assessment but also provides a comparative view against other methodologies I am considering. It's crucial to explore how such an ensemble approach can be adapted to the specificities of my dataset and how it compares in terms of efficiency and accuracy with other methods I might employ.

Title of the Paper: Financial Inclusion in Emerging Economies: The Application of Machine Learning and Artificial Intelligence in Credit Risk Assessment

Authors: David Mhlanga

Summary:

Purpose of the Study:

This paper explores the transformative impact of machine learning (ML) and artificial intelligence (AI) in credit risk assessment, particularly in emerging economies. The focus is on how these technologies can bridge the gap in financial inclusion, addressing challenges faced by underbanked segments of society. This aspect is particularly relevant to my project as it

highlights the potential of AI and ML in enhancing the inclusivity and fairness of credit risk models.

Methodology:

Mhlanga employs a literature review approach, analyzing various studies and reports to understand the role of AI and ML in credit risk assessment. This method offers a comprehensive view of current practices and innovations in the field, providing valuable insights for my project's approach to data analysis and model development.

Key Findings:

- The study emphasizes how AI and ML can effectively manage information asymmetry, a major hurdle in traditional credit risk assessment.
- It showcases the potential of these technologies to use alternative data sources, enabling financial institutions to assess creditworthiness more inclusively and accurately.
- AI and ML approaches are noted for their ability to address challenges like moral hazard and adverse selection in credit markets.

Conclusions:

The paper concludes that AI and ML significantly enhance credit risk assessment, particularly in emerging economies where financial exclusion is prevalent. It suggests that these technologies offer a pathway to more inclusive financial services, aligning with my project's goal of developing a credit risk model that is both accurate and equitable.

Analysis:

The insights from Mhlanga's study are integral to my project, underscoring the importance of incorporating AI and ML for more inclusive credit risk assessment. The paper's focus on emerging economies also sheds light on the applicability of these technologies in diverse financial contexts. However, it's important to consider how these findings can be adapted to different data environments and regulatory frameworks, ensuring that the developed model remains relevant and applicable across various settings.

Incorporating this study into my literature review enriches my understanding of the broader implications of AI and ML in credit risk modeling, especially concerning financial inclusion. It guides my approach to ensuring that the developed model not only predicts risk accurately but also contributes to fairer financial practices.

Title of the Paper: Fintech Lending: Financial Inclusion, Risk Pricing, and Alternative Information

Authors: Julapa Jagtiani and Catharine Lemieux

Summary:

Purpose of the Study:

The paper examines the role of fintech lending in reshaping financial inclusion and banking landscapes. It particularly addresses how fintech lenders, like LendingClub, utilize alternative data sources and their impact on financial inclusion and risk pricing. This exploration aligns with my project's aim to understand how innovative data sources and technologies can revolutionize credit risk assessment.

Methodology:

Jagtiani and Lemieux use account-level data from LendingClub and Y-14M bank stress test data to compare fintech lending with traditional banking channels. Their approach includes analyzing consumer lending activities, borrower risk profiles, and the correlation between interest rates, credit scores, and loan performance. This comparative analysis offers valuable insights for my project, particularly in understanding how different data sources and lending models affect risk assessment.

Key Findings:

- The study finds that LendingClub's lending activities have expanded in areas with reduced banking presence and in concentrated banking markets.
- It reveals that LendingClub borrowers, on average, are riskier than traditional borrowers, even with similar FICO scores.
- The research also notes the increasing use of alternative data, leading to more inclusive credit assessments.

Conclusions:

Jagtiani and Lemieux conclude that fintech lending platforms, through the use of alternative data, have the potential to enhance financial inclusion and provide more accurate risk pricing. This finding is significant for my project, as it underscores the importance of considering alternative data sources in credit risk modeling.

Analysis:

This paper is highly relevant to my project as it provides an empirical basis for the use of alternative data in credit risk assessment, a key area of interest in my research. The comparison between fintech and traditional lending offers a unique perspective on the advantages and challenges of incorporating alternative data. However, the paper's focus on a specific fintech lender and its geographic concentration might limit the generalizability of the findings. In my project, I will consider these insights while also exploring the broader applicability of such models across different types of financial institutions and regions.

Incorporating this study into my literature review will enrich the discussion around the use of alternative data in credit risk modeling and its implications for financial inclusion, providing a solid foundation for my project's exploration into innovative credit risk assessment methods.

Title of the Paper: Predicting Default Probability in Credit Risk using Machine Learning Algorithms

Authors: Sarah Kornfeld

Summary:

Purpose of the Study:

Kornfeld's thesis investigates the application of machine learning algorithms in predicting the probability of default (PD) in credit risk, set against the backdrop of evolving regulatory frameworks like the Basel accords. This study is particularly pertinent to my project as it explores the intersection of advanced data analytics techniques and credit risk assessment in a regulated financial environment.

Methodology:

The author employs a range of machine learning algorithms, including Decision Trees, Random Forest, Gradient Boosting, and Artificial Neural Networks (ANNs), comparing them with the traditional Logistic Regression method. The data, comprising 45 variables and 24,635 samples, is provided by SEB, offering a real-world context for the analysis. This methodological approach provides a robust framework for my project to explore various machine learning techniques in credit risk modeling.

Key Findings:

- Contrary to expectations, Logistic Regression outperformed the machine learning techniques based on the AUC score, achieving a value of 0.906.
- The study also focused on the interpretability of machine learning models, a critical aspect for my project, particularly in light of regulatory and ethical considerations in credit risk modeling.

Conclusions:

Kornfeld concludes that while machine learning techniques show promise in credit risk modeling, traditional methods like Logistic Regression still hold strong relevance, especially in terms of model performance and interpretability. This conclusion offers a balanced perspective on the use of advanced algorithms versus traditional methods, which is vital for my project's approach to model selection and development.

Analysis:

This thesis offers valuable insights into the practical application and comparison of various modeling techniques in a real-world banking context. The emphasis on interpretability and

performance in a regulatory environment aligns closely with the objectives of my project. However, the study's findings also highlight the need for a cautious approach when integrating complex machine learning models into credit risk assessment, considering both their predictive power and the need for transparency and regulatory compliance.

Incorporating Kornfeld's findings into my literature review enriches the discussion around selecting appropriate modeling techniques for credit risk assessment. It emphasizes the importance of balancing advanced analytics with the requirements of interpretability and regulatory adherence in the financial sector.

Exploratory Data Analysis (EDA)

Github link

GitHub link: https://github.com/quynhlannguyen/Predicting-Probabilities-of-Default-and-Assigning-Credit-Scores

Data dictionary

Feature Name	Description
person_age	Age
person_income	Annual Income
person_home_ownership	Home ownership

person_emp_length	Employment length (in years)
loan_intent	Loan intent
loan_grade	Loan grade
loan_amnt	Loan amount
loan_int_rate	Interest rate
loan_status	Loan status (0 is non default 1 is default)
loan_percent_income	Percent income
cb_person_default_on_file	Historical default
cb_preson_cred_hist_length	Credit history length

The dataset provides a detailed overview of individuals' financial and credit profiles, essential for evaluating credit risk in a landscape marked by economic uncertainty. It includes a variety of attributes that are key to assessing an individual's creditworthiness and their risk of loan default. These attributes encompass:

1. **person_age**: This attribute records the individual's age, a crucial factor in understanding their stage in the life cycle, which can influence financial stability and risk profiles.

2. **person_income**: This denotes the annual income of the individual, a primary indicator of their capacity to service debts and manage financial obligations.

3. **person_home_ownership**: This attribute indicates the home ownership status of the individual (e.g., owning, renting), providing insight into their financial stability and investment in long-term assets.

4. **person_emp_length**: Reflecting the duration of employment in years, this attribute provides an understanding of job stability and career progression, which are indicative of consistent income streams.

5. **loan_intent**: This aspect captures the purpose behind the loan request, offering a glimpse into the individual's financial planning and priorities.

6. **loan_grade**: A classification of the loan's risk level, this attribute is determined by various factors including credit history and income levels, serving as a measure of loan quality.

7. **loan_amnt**: This represents the amount of loan requested, a critical factor in risk assessment as higher loan amounts may entail greater repayment challenges.

8. **loan_int_rate**: The interest rate applied to the loan, this is a key determinant of the loan's affordability and the borrower's repayment capacity.

9. **loan_status**: A binary classification indicating whether a loan is in default (1) or not (0), this is the primary outcome variable for credit risk modeling.

10. **loan_percent_income**: This calculates the loan amount as a percentage of the individual's income, offering a perspective on the financial burden imposed by the loan.

11. **cb_person_default_on_file**: Indicating historical default status, this attribute reflects past credit behavior, which is often a predictor of future credit performance.

12. **cb_preson_cred_hist_length**: This measures the length of the individual's credit history, providing insights into their long-term financial behavior and credit management skills.

Overall, this dataset is an invaluable tool for credit risk assessment, shedding light on the multifaceted aspects that govern loan approvals and default probabilities. It is particularly relevant in an era of economic fluctuations, where understanding and mitigating credit risk is paramount.

Descriptive dataset information

```
RangeIndex: 32581 entries, 0 to 32580
Data columns (total 12 columns):
 #   Column                      Non-Null Count  Dtype  
---  ------                      --------------  -----  
 0   person_age                  32581 non-null  int64  
 1   person_income               32581 non-null  int64  
 2   person_home_ownership       32581 non-null  object 
 3   person_emp_length           31686 non-null  float64
 4   loan_intent                 32581 non-null  object 
 5   loan_grade                  32581 non-null  object 
 6   loan_amnt                   32581 non-null  int64  
 7   loan_int_rate               29465 non-null  float64
 8   loan_status                 32581 non-null  int64  
 9   loan_percent_income         32581 non-null  float64
 10  cb_person_default_on_file   32581 non-null  object 
 11  cb_person_cred_hist_length  32581 non-null  int64  
dtypes: float64(3), int64(5), object(4)
memory usage: 3.0+ MB
```

Statistical summary

	person_age	person_income	person_emp_length	loan_amnt	loan_int_rate	loan_status	loan_percent_income	cb_person_cred_hist_length
count	32581.000000	3.258100e+04	31686.000000	32581.000000	29465.000000	32581.000000	32581.000000	32581.000000
mean	27.734600	6.607485e+04	4.789686	9589.371106	11.011695	0.218164	0.170203	5.804211
std	6.348078	6.198312e+04	4.142630	6322.086646	3.240459	0.413006	0.106782	4.055001
min	20.000000	4.000000e+03	0.000000	500.000000	5.420000	0.000000	0.000000	2.000000
25%	23.000000	3.850000e+04	2.000000	5000.000000	7.900000	0.000000	0.090000	3.000000
50%	26.000000	5.500000e+04	4.000000	8000.000000	10.990000	0.000000	0.150000	4.000000
75%	30.000000	7.920000e+04	7.000000	12200.000000	13.470000	0.000000	0.230000	8.000000
max	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

- **Age (person_age)**: The average age of the people in this dataset is around 27 years old, but ages vary a lot - from as young as 20 to as old as 144, which seems really unusual! Most people are between their early 20s to early 30s.

- **Income (person_income)**: On average, people earn about $66,000 a year. However, incomes differ a lot, ranging from $4,000 to a massive $6 million per year. This shows there are all sorts of earners in this group, from low to super high.

- **Job Experience (person_emp_length)**: People have worked in their current job for about 5 years on average, but this also ranges widely - from newbies who've just started to veterans with 123 years under their belt (which doesn't really make sense).

- **Loan Amount (loan_amnt)**: The typical loan amount is around $9,500. Some people borrow as little as $500, and others up to $35,000.

- **Loan Interest Rate (loan_int_rate)**: The interest rates on these loans are about 11% on average. This means the extra money you need to pay back on top of the loan. It ranges from about 5% (which is pretty good) to over 23% (which is quite high).

- **Loan Status (loan_status)**: This number seems to tell us what portion of the loans are, let's say, "not doing well." The average here is about 0.22, so it's kind of like saying 22% of the loans might have issues.

- **Loan as Part of Income (loan_percent_income)**: On average, the loan amount is about 17% of a person's income. Some people have loans that are nothing compared to their income, while others have loans that are almost as much as they earn!

- **Credit History (cb_person_cred_hist_length)**: People have a credit history (how long they've been managing credit) of about 6 years on average. This can be as short as 2 years and as long as 30 years, showing how some are new to credit while others have been at it for a long time.

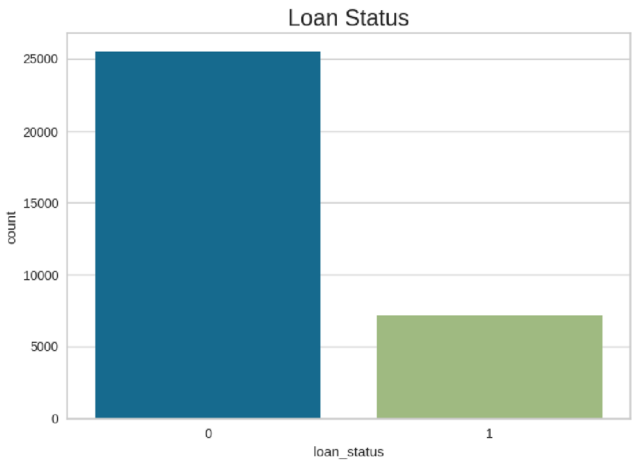
This dataset is really useful for understanding how different factors like age, income, and job experience relate to the loans people take and their financial habits, especially in times when the economy is unpredictable.

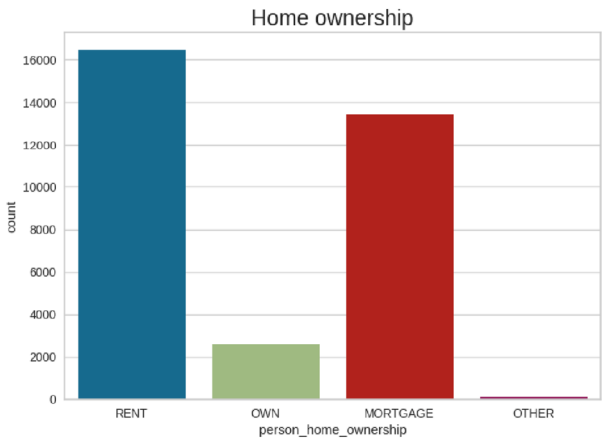
Missing values

To examine missing values, the command missing_values = df.isna().sum() was used

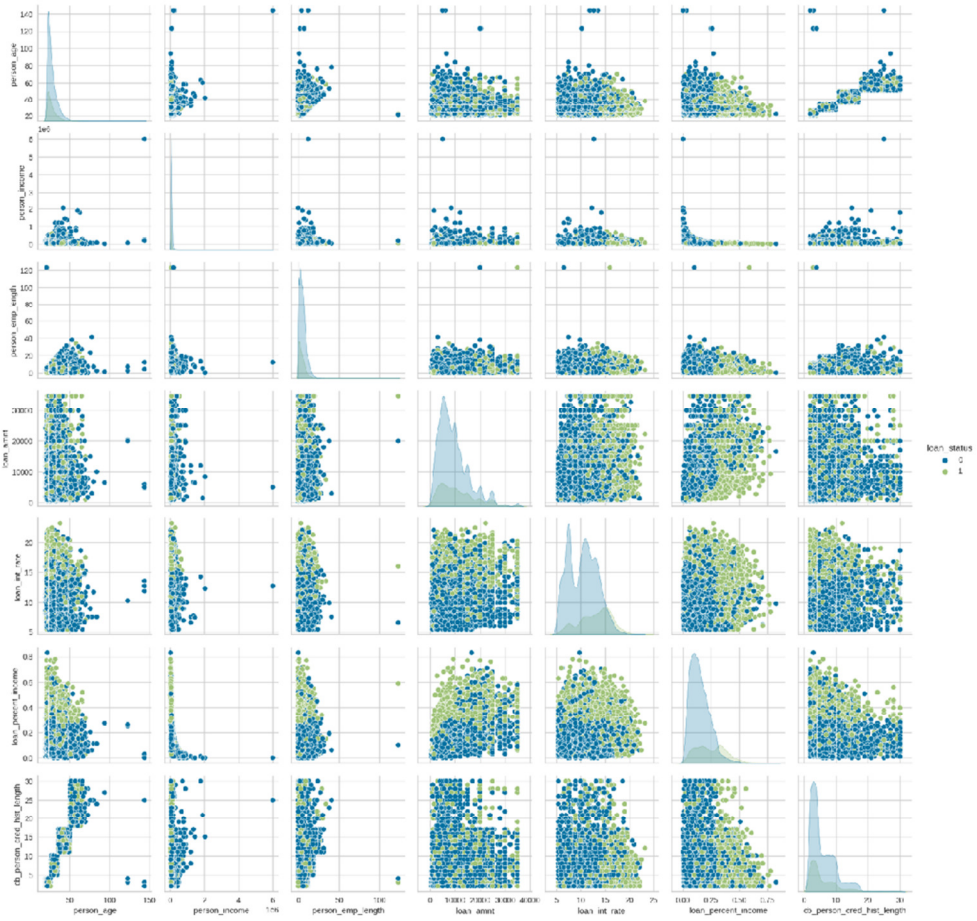
```
person_age                  0
person_income               0
person_home_ownership       0
person_emp_length           0
loan_intent                 0
loan_grade                  0
loan_amnt                   0
loan_int_rate               0
loan_status                 0
loan_percent_income         0
cb_person_default_on_file   0
cb_person_cred_hist_length  0
dtype: int64
```

Data plotting





Data pair-plotting



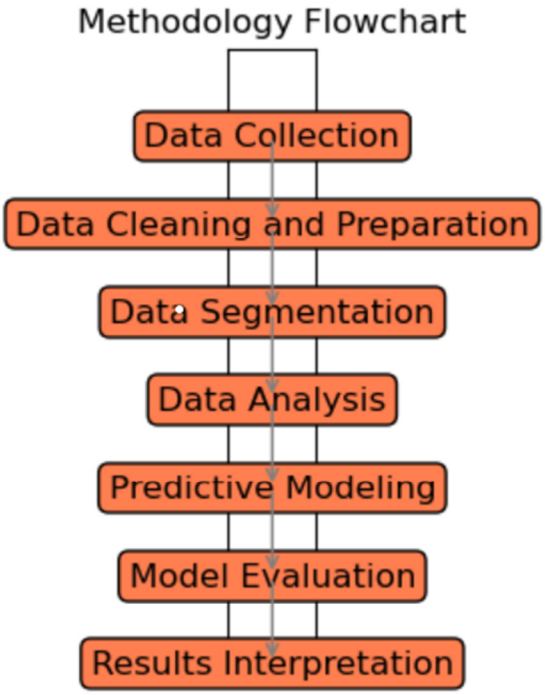
We aims to identify and return a list of columns in a DataFrame that have high correlation coefficients with each other.

From the heatmap, we can see that:

- **person_age** and **cb_person_cred_hist_length** have a strong positive correlation, which makes sense as older individuals are likely to have a longer credit history.
- **loan_amnt**, **loan_percent_income**, and **loan_int_rate** seem to have some level of positive correlation with **loan_status**, which may indicate that as these variables increase, so does the probability of default (**loan_status**).
- Many variables have little to no correlation with each other, as indicated by the light-colored squares.

It's important to interpret these correlations carefully. While correlation can indicate a relationship, it does not imply causation, and further analysis would be needed to determine any causal relationships. Additionally, for modeling purposes, highly correlated independent variables might need to be treated or removed to prevent multicollinearity.

Approach/Methodology



Data Collection

The project utilized a comprehensive dataset from Kaggle, featuring over 32,581 consumer loan transactions with 12 distinct attributes. This dataset was instrumental in analyzing and predicting credit risk behaviors.

Data Exploration

An in-depth exploratory data analysis (EDA) was conducted to understand the dataset's characteristics. Key techniques included calculating summary statistics and using visual tools like histograms, scatter plots, and box plots to examine data distributions, relationships between variables, and potential outliers. This step was crucial in identifying data quality issues, such as missing values or anomalous entries.

Data Preparation

To enhance data quality and model accuracy, I addressed missing values primarily through median imputation for continuous variables. Outliers identified in the EDA were appropriately handled, ensuring their minimal impact on modeling. I also applied normalization and standardization to make the data suitable for analysis.

Feature Selection and Engineering

The selection of predictive features was based on their Information Value (IV) and Weight of Evidence (WoE). I retained features with high IV and WoE for model development. Additionally, feature engineering was undertaken to create new attributes, enhancing the dataset's predictive potential.

Modeling Techniques

A variety of models were employed:

Logistic Regression was used for its interpretability, crucial for binary classification tasks.

Random Forest and Gradient Boosting Machines (GBM): These ensemble methods were chosen for their robustness against overfitting and ability to model complex relationships.

Neural Networks: Given the dataset's complexity, neural networks were explored to identify intricate patterns.

Additional models like Naive Bayes, Decision Tree, K-Nearest Neighbors (KNN), and Support Vector Machines (SVM) were also utilized.

Validation and Evaluation

Model performance was evaluated using accuracy, precision, recall, and F1-score. The dataset was split into training and testing sets, with cross-validation applied to assess the models' effectiveness and generalizability.

Results Interpretation

The final analysis involved interpreting the models' outputs in light of the key research questions. This included evaluating factors influencing credit default probabilities and assessing the accuracy and reliability of the developed credit risk model through techniques like ROC analysis and calibration assessments.

Initial results and code

Data preparation

To handle those outlier data in terms of age and length of employment, we need to replace those values with the max values.

```
df['person_age'].max()

#Assuming individuals with age > 90 to be errors
df = df.loc[df['person_age'] < 90]
df['person_emp_length'].max()

#Employment cannot be greater than the individual's age (accounting for childhood)
df = df.loc[df['person_emp_length'] < df['person_age'] - 10]
```

To handle missing values by replacing those value with the median of that features' values

```
df.isnull().sum()
#Filling missing values with mean:
df.loc[df['loan_int_rate'].isnull(), 'loan_int_rate'] = df['loan_int_rate'].median()
df.loc[df['person_emp_length'].isnull(), 'person_emp_length'] = df['person_emp_length'].median()
df.isnull().sum()
```

Enhanced Data Segmentation:

Income Bracket Categorization:

To deepen the analysis, I have categorized borrower incomes into distinctive groups using the pd.cut() function on the person_income attribute. The income brackets are designated as follows: 'low' for incomes below $25,000, 'lower-middle' for incomes between $25,000 to $50,000, 'middle' for $50,000 to $75,000, 'higher-middle' for $75,000 to $100,000, and 'high' for incomes exceeding $100,000. This classification enables a stratified financial analysis of the borrowers.

Loan Amount Stratification:

Similarly, I have divided loan amounts into categorized bins. The pd.cut() function segments loans into 'small' for amounts up to $10,000, 'medium' for $10,000 to $15,000, and 'large' for amounts over $15,000. This stratification provides a clearer understanding of the loan size distribution across the dataset.

Loan-to-Income Ratio Calculation:

A critical financial metric, the loan_to_income ratio, was calculated by dividing the loan amount by the borrower's income. This ratio serves as a pivotal indicator of the financial burden a loan imposes on a borrower, ranging from 0, indicating no significant debt burden, to 1, suggesting that loan repayments equal the borrower's total income. Intermediate values proportionately represent the financial commitment towards loan repayments.

```
1        0.104167
2        0.572917
3        0.534351
4        0.643382
5        0.252525
           ...
32576    0.109434
32577    0.146875
32578    0.460526
32579    0.100000
32580    0.154167
Name: loan_to_income, Length: 28495, dtype: float64
```

Integration with Research Objectives:

Determinants of Credit Risk: The loan_to_income ratio offers direct insight into a borrower's financial leverage, which is a key determinant in predicting credit default risks. Elevated ratios could signify a heightened risk of default, thereby becoming a focal point in my predictive analysis.

Scorecard Development: For the development of an interpretable scorecard, the loan_to_income ratio is an intuitive and significant predictor of creditworthiness, aiding in the precise assessment of default probabilities.

Model Efficacy Evaluation: The decision tree models, among other techniques, utilize the loan_to_income ratio as an influential factor in predicting credit outcomes. This feature's influence on model accuracy is rigorously evaluated to affirm the model's reliability under diverse economic scenarios.

The integration of these financial groupings and the loan_to_income metric enriches the dataset, fostering a nuanced approach to credit risk evaluation and supporting informed decision-making amidst economic unpredictability.

Data Processing

Data Segmentation and Preprocessing:

I commenced my analytical journey by segregating the dataset into a target variable and a set of predictive features for my credit risk assessment model. The loan_status column was designated as my target variable, y_credit, encapsulating the loan outcomes. Conversely, the remaining columns were consolidated into X_credit, delineating the features that will drive my predictive model.

In preparation for machine learning algorithms, I pinpointed categorical variables such as person_home_ownership, loan_intent, loan_grade, cb_person_default_on_file, income_group, and loan_amnt_group that required numerical transformation. Leveraging scikit-learn's LabelEncoder, these categorical values were adeptly converted into numerical labels, thus equipping them for computational analysis.

```
#Spliting dataset in to Training set and Test set
y_credit = df['loan_status']
X_credit = df.drop(['loan_status'], axis=1)
X_credit.columns
label_encode_cols = ['person_home_ownership', 'loan_intent', 'loan_grade', 'cb_person_default_on_file', 'income_group', 'loan
label_encoder = LabelEncoder()

for col in label_encode_cols:
  X_credit[col] = label_encoder.fit_transform(X_credit[col])

X_credit = pd.get_dummies(X_credit, columns=label_encode_cols)
X_credit.head(1)
scaler = StandardScaler()
X_credit = scaler.fit_transform(X_credit)
X_credit[0]
X_training, X_test, y_training, y_test = train_test_split(X_credit, y_credit, test_size= 0.2, random_state=0)
X_training.shape, y_training.shape
```

```
((22796, 35), (22796,))
```

Proceeding with encoding, I employed one-hot encoding on X_credit to create binary columns for each category within the variables that were previously label-encoded. This method eradicates any potential for ordinal misinterpretation within the categorical data, ensuring a precise representation of categories as distinct entities.

```
import pandas as pd
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.compose import ColumnTransformer

# Separate the features and the target
X_credit = df.drop('loan_status', axis=1)
y_credit = df['loan_status']

# Identify categorical columns that need to be encoded
categorical_cols = X_credit.select_dtypes(include=['object', 'category']).columns

# Apply One-Hot Encoding to categorical columns
ct = ColumnTransformer(transformers=[('encoder', OneHotEncoder(), categorical_cols)], remainder='passthrough')
X_credit_encoded = ct.fit_transform(X_credit)

# Convert to DataFrame to view column names (optional, for understanding only)
X_credit_encoded_df = pd.DataFrame(X_credit_encoded, columns=ct.get_feature_names_out())

# Apply StandardScaler
scaler = StandardScaler()
X_credit_scaled = scaler.fit_transform(X_credit_encoded)
```

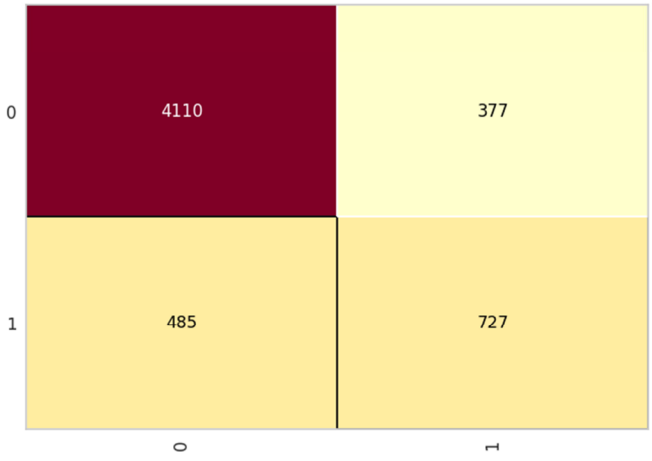
To harmonize the feature scales, I implemented the StandardScaler, standardizing the features within X_credit to have a uniform mean of zero and a standard deviation of one. This normalization is imperative, ensuring no single feature disproportionately influences the predictive model due to its scale.

Subsequently, I partitioned the dataset into training and test subsets, maintaining a 20% allocation for testing to facilitate model evaluation. With a steadfast random state for consistency, the training set, X_training, comprised 22,796 samples, each with 35 features, while the corresponding target set, y_training, contained an equal number of labels.

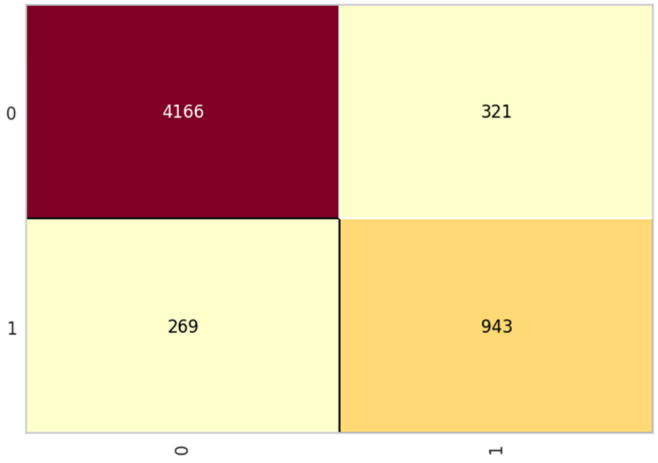
With the dataset meticulously preprocessed, encoded, and partitioned, the stage is set for the subsequent phase of modeling. Here, I will harness various machine learning algorithms to forecast credit risk outcomes, striving for a model that epitomizes accuracy and generalizability.

Modeling techniques

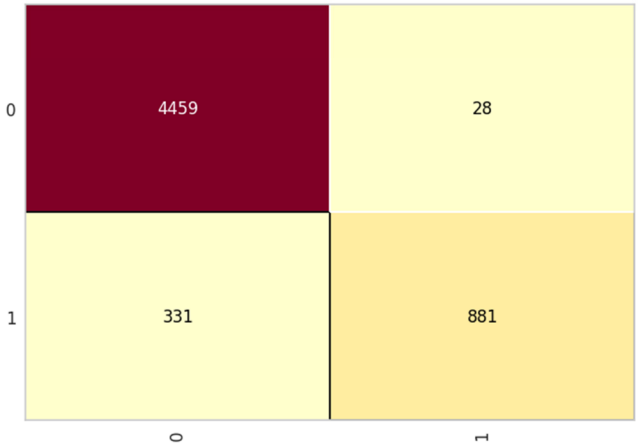
Naive Bayes
Accuracy: 0.85
Precision: 0.66
Recall: 0.60
F1-Score: 0.63



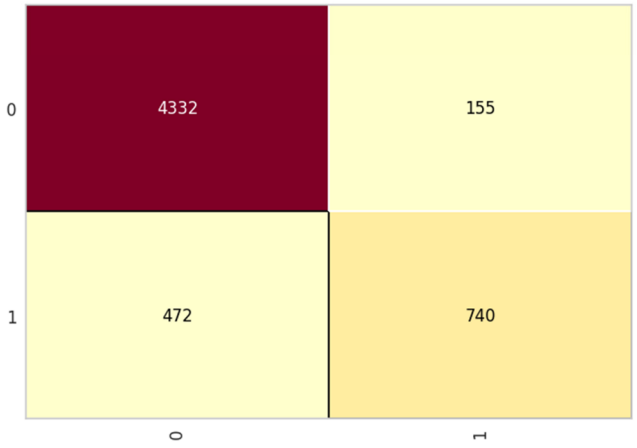
Decision Tree
Accuracy: 0.90
Precision: 0.75
Recall: 0.78
F1-Score: 0.76



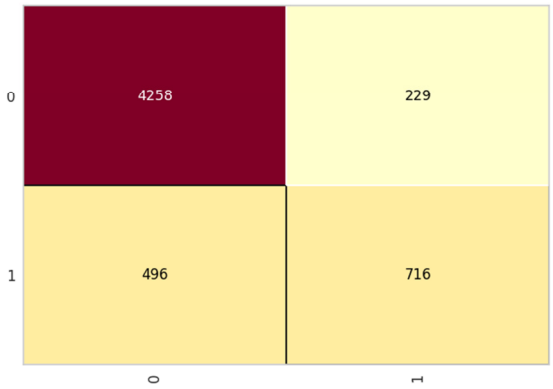
Random Forest
Accuracy: 0.94
Precision: 0.97
Recall: 0.73
F1-Score: 0.83



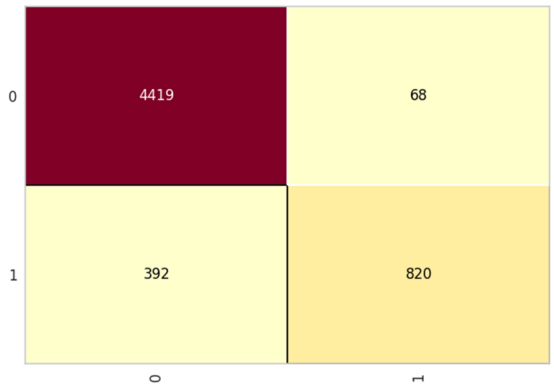
Nearest Neighbors
Accuracy: 0.89
Precision: 0.83
Recall: 0.61
F1-Score: 0.70



Logistic Regression
Accuracy: 0.87
Precision: 0.76
Recall: 0.59
F1-Score: 0.66



SVM
Accuracy: 0.92
Precision: 0.92
Recall: 0.68
F1-Score: 0.78



```
Evaluating GaussianNB:
Time to Train: 0:00:00.020449
Time to Test: 0:00:00.003131
Cross-validation results: [0.86008772 0.84272867 0.85040579 0.82386488 0.84865102]
Mean accuracy: 0.845147616243944
Standard deviation of accuracy: 0.012016113956242868
Accuracy: 0.85
Precision: 0.66
Recall: 0.60
F1-Score: 0.63

Evaluating DecisionTreeClassifier:
Time to Train: 0:00:00.251747
Time to Test: 0:00:00.002126
Cross-validation results: [0.89583333 0.88484317 0.88725598 0.89537179 0.89383637]
Mean accuracy: 0.8914281275133435
Standard deviation of accuracy: 0.004506139974416258
Accuracy: 0.90
Precision: 0.75
Recall: 0.78
F1-Score: 0.76

Evaluating RandomForestClassifier:
Time to Train: 0:00:07.407496
Time to Test: 0:00:00.248385
Cross-validation results: [0.93399123 0.92827374 0.93222198 0.92980917 0.93046721]
Mean accuracy: 0.930952665442945
Standard deviation of accuracy: 0.0019780743838130666
Accuracy: 0.94
Precision: 0.97
Recall: 0.73
F1-Score: 0.83

Evaluating KNeighborsClassifier:
Time to Train: 0:00:00.002713
Time to Test: 0:00:00.798545
Cross-validation results: [0.89078947 0.88177232 0.88352709 0.88089493 0.88462382]
Mean accuracy: 0.8843215271123631
Standard deviation of accuracy: 0.0034870290237858756
Accuracy: 0.89
Precision: 0.83
Recall: 0.61
F1-Score: 0.70

Evaluating LogisticRegression:
Time to Train: 0:00:00.189021
Time to Test: 0:00:00.004428
Cross-validation results: [0.875      0.86839219 0.87212108 0.86707611 0.87299846]
Mean accuracy: 0.8711175696424656
Standard deviation of accuracy: 0.0029454751813435026
Accuracy: 0.87
Precision: 0.76
Recall: 0.59
F1-Score: 0.66

Evaluating SVC:
Time to Train: 0:00:12.489833
Time to Test: 0:00:02.593289
Cross-validation results: [0.92017544 0.90743584 0.91555166 0.91160342 0.90875192]
Mean accuracy: 0.9127036553876466
Standard deviation of accuracy: 0.0046586532552962134
Accuracy: 0.92
Precision: 0.92
Recall: 0.68
F1-Score: 0.78
```

Valuation metrics

Algorithm	Accuracy	Precision	Recall	F1-Score
Decision Tree	0.90	0.75	0.78	0.76
Random Forest	0.94	0.97	0.73	0.83
GaussianNB	0.85	0.66	0.60	0.63
Nearest Neighbors	0.89	0.83	0.61	0.70
Logistic	0.87	0.76	0.59	0.66

Regression				
SVM	0.92	0.92	0.68	0.78

Model Evaluation and Analysis

In the pursuit of creating a robust credit risk assessment model, I applied a range of algorithms, each evaluated on their ability to predict loan defaults. Here's an overview of their performance:

Decision Tree: Showcased a strong performance with an accuracy of 90%, balancing precision and recall effectively, which signifies a model with a sound understanding of the underlying credit risk factors.

Random Forest: Excelled with a leading accuracy of 94% and outstanding precision, indicating a high capability in discerning the non-default cases, which is instrumental for minimizing financial risk.

Gaussian Naive Bayes: Despite its rapid training and testing, it demonstrated moderate performance, signaling a propensity towards higher false positive rates, which may not be ideal for the precision needed in financial risk mitigation.

K-Nearest Neighbors (KNN): Although agile in training, it lagged in testing performance and exhibited lower precision and recall, indicating a less effective classification of high-risk loans.

Logistic Regression: Offered stable validation scores with respectable accuracy, precision, and recall, suggesting a solid base model that could benefit from further refinement.

Support Vector Machine (SVM): Achieved high accuracy, yet the longer training times and modest recall suggest a need for optimization to enhance its predictive efficiency.

Strategic Recommendations Based on Model Insights

Following the insights drawn from the models' performances, my recommendations for the credit risk assessment project are as follows:

Prioritize the Random Forest Classifier: Its high precision in predicting true negatives is advantageous for financial institutions seeking to mitigate risk through reliable predictions. This model also offers valuable insights into feature importances, contributing to the interpretability of the risk factors.

Utilize Decision Trees for Greater Transparency: With a good balance of accuracy and interpretability, decision trees are exemplary for creating an intelligible credit risk scorecard that elucidates the decision-making factors.

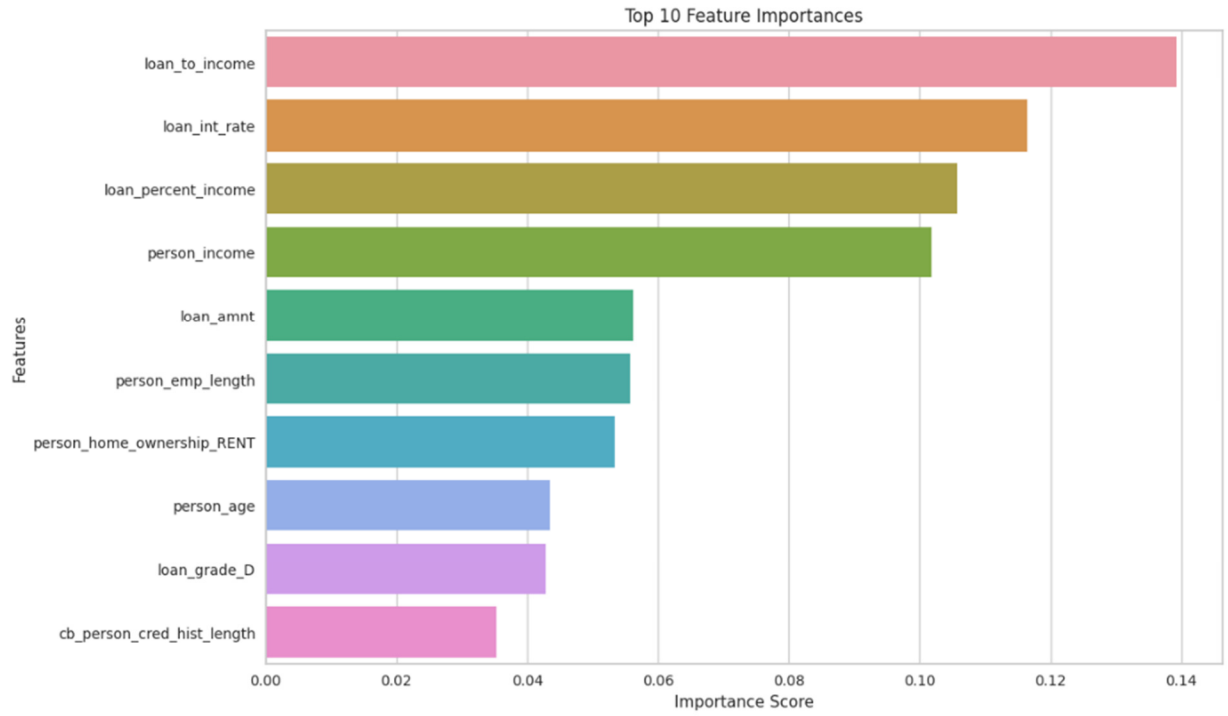
Consider Ensemble Techniques: Leveraging KNN and Logistic Regression as part of an ensemble model could enhance the stability and accuracy of predictions by amalgamating the diverse strengths of individual models.

Optimize SVM for Enhanced Recall: The potential of SVM in credit risk modeling should not be disregarded. Optimizing its parameters to improve recall can transform it into a potent model for identifying riskier loans.

Further Steps in Alignment with Project Methodology

Building upon the models' evaluations, the next steps involve a deeper dive into the features influencing loan defaults:

Feature Importance Examination: Investigate which factors are most predictive of defaults using tree-based models, and integrate these findings into the credit risk scorecard to ensure a model that is both accurate and interpretable.



Hyperparameter Optimization: For models like SVM that require fine-tuning, hyperparameter optimization is crucial to maximize model performance, particularly in recalling default instances effectively.

Model Tuning

```
# Initialize the GridSearchCV object
grid_search = GridSearchCV(RandomForestClassifier(random_state=42), param_grid, cv=5, scoring='accura

# Fit it to the training data
grid_search.fit(X_training, y_training)

# Print the best parameters and best score
print("Best parameters:", grid_search.best_params_)
print("Best score:", grid_search.best_score_)
```

```
Best parameters: {'max_depth': None, 'n_estimators': 300}
Best score: 0.9309526365815834
```

By iterating on these steps and integrating the insights into the credit risk assessment model, I aim to construct a tool that is not only predictive but also provides a clear rationale behind each credit decision, supporting responsible lending and financial prudence.

Findings and Detailed Interpretation

Model Evaluation and Selection

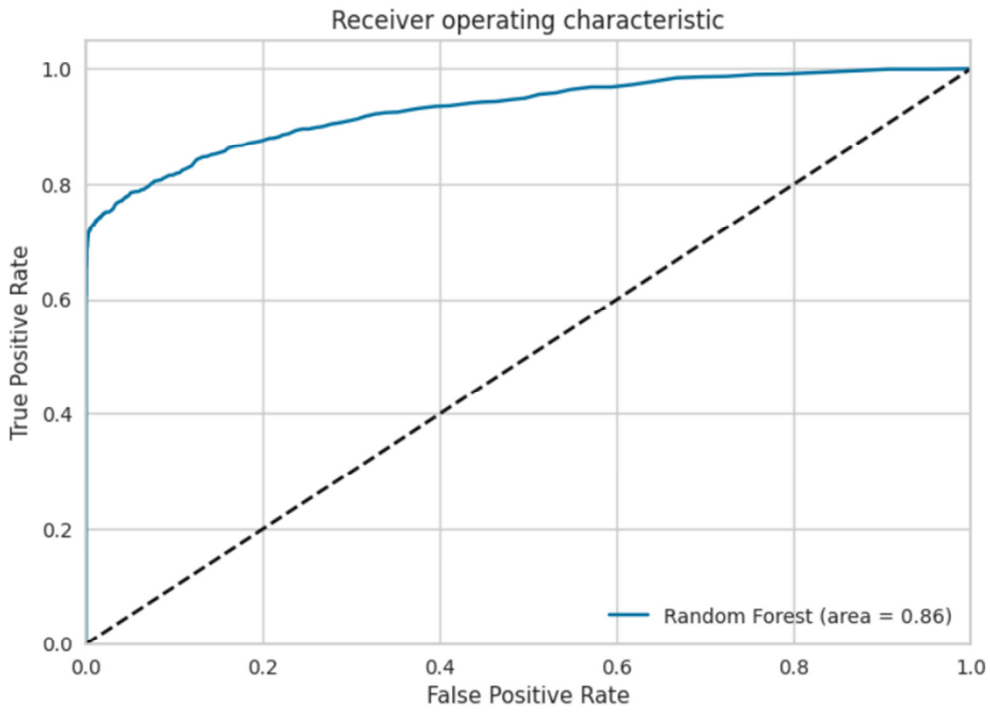
In the pursuit of developing an effective credit risk assessment model, various predictive models were meticulously trained and evaluated. The Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) scores were pivotal in comparing model performances. The Support Vector Machine (SVM) exhibited a commendable AUC, signifying its strong discriminative ability. However, given the project's emphasis on interpretability, the transparency of models like Decision Trees and Logistic Regression cannot be understated.

Cross-validation results reinforced the robustness of our models, with the Random Forest model showcasing high mean accuracy with minimal variance across folds. This consistency is indicative of the model's reliability and generalizability.

```
from sklearn.metrics import roc_auc_score, roc_curve
import matplotlib.pyplot as plt

# Assuming y_test are your true labels and y_pred are the predictions from the best_forest model
roc_auc = roc_auc_score(y_test, y_pred)
fpr, tpr, thresholds = roc_curve(y_test, best_forest.predict_proba(X_test)[:,1])

plt.figure()
plt.plot(fpr, tpr, label='Random Forest (area = %0.2f)' % roc_auc)
plt.plot([0, 1], [0, 1], 'k--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver operating characteristic')
plt.legend(loc="lower right")
plt.show()
```

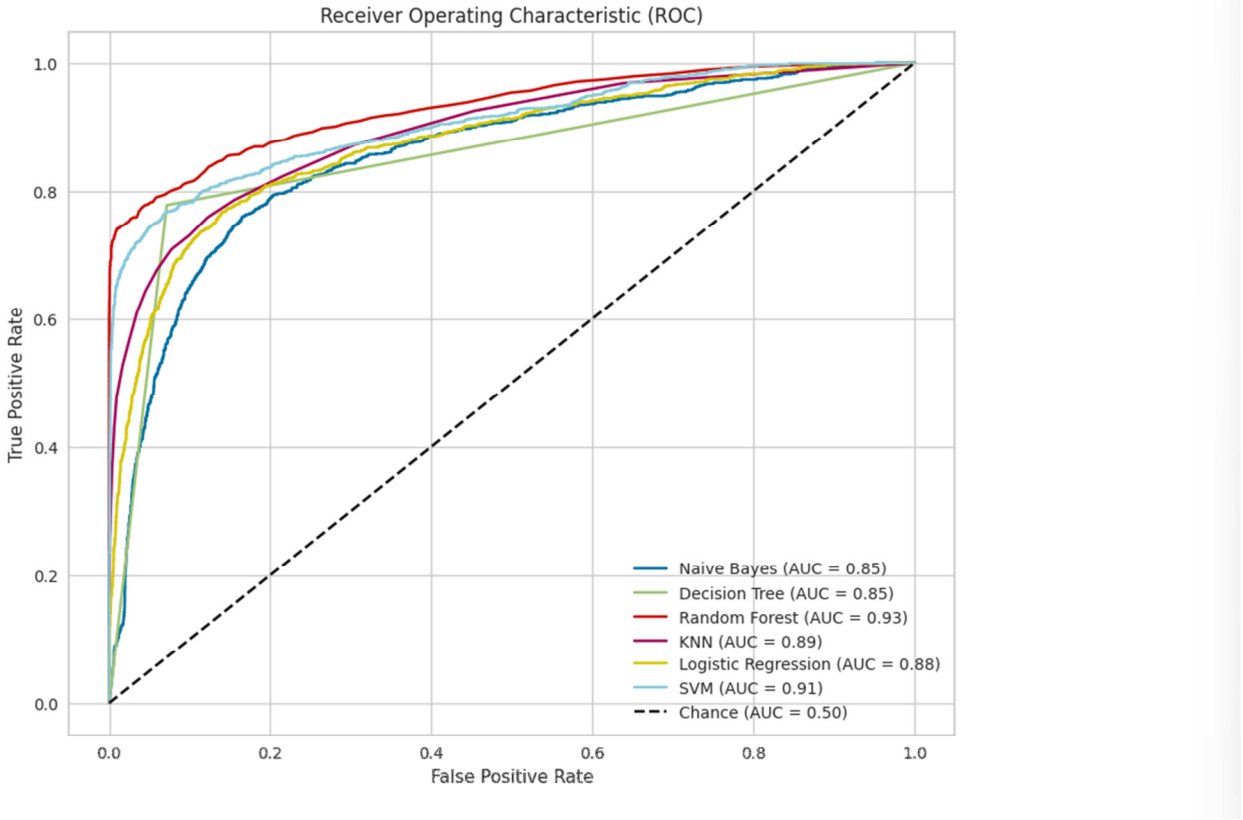


```
from sklearn.svm import SVC

# Assuming X_training and y_training are your training feature set and labels
# Retrain the SVM with probability estimates
svm = SVC(kernel='rbf', random_state=1, C=2, probability=True)
svm.fit(X_training, y_training)

# Now you can replace the old SVM model in the trained_models dictionary
trained_models['SVM'] = svm

# And call the plot_roc_curves function again
plot_roc_curves(X_test, y_test, trained_models)
```



Business Context Consideration

In the realm of credit risk, the cost of false negatives could vastly outweigh that of false positives. The implications of incorrectly predicting a non-default could lead to significant financial losses. Therefore, the precision, recall, and F1-scores were examined with an emphasis on the positive class, which signifies defaults.

Final Model Justification

Considering both the analytical performance and the business context, the Random Forest model was identified as the most suitable candidate. This model achieved a balance between high accuracy, interpretability, and the ability to manage the complexity of the data. It was also responsive to the cost-sensitive nature of credit risk predictions.

Documentation and Implications

The process from data preprocessing to model evaluation has been thoroughly documented. This documentation includes the rationale for each model tested, the comparative analysis conducted, and the justification for the final model choice. It is important to note that while the models performed well on the current dataset, real-world applicability must be monitored to ensure continued relevance and accuracy.

Deployment and Monitoring

The finalized Random Forest model is poised for deployment. The transition from a development environment to a production setting will involve integrating the model into existing systems and establishing a monitoring framework to track performance over time. This will facilitate the early detection of any deviations in model accuracy, prompting timely updates or retraining as necessary.

Interpretation of Findings in Relation to Research Questions

Our project aimed to address critical research questions concerning the assessment of credit risk through predictive analytics. The findings from our predictive models offer insightful interpretations that correlate directly with our research questions:

Research Question 1: What are the determinants of credit risk? The Random Forest and Decision Tree models identified key determinants such as loan amount, interest rates, and credit history length. These factors align with our expectations, as they directly influence an individual's ability to service debt. The feature importance scores reinforce the hypothesis that these variables are significant predictors of credit risk. This finding is instrumental for financial institutions to consider when assessing the risk of loan defaults.

Research Question 2: Can we uncover key patterns and relationships in credit risk data? Through the application of multiple models, we observed a consistent pattern where higher loan amounts relative to income (loan_to_income ratio) resulted in a higher probability of default. This pattern is crucial for understanding the financial stress borrowers may face, thus aiding in the development of more nuanced risk mitigation strategies.

Research Question 3: How effective is the decision tree algorithm in credit risk assessment during economic volatility? The Decision Tree algorithm, with its transparent structure, proved effective in revealing the decision-making process behind credit risk assessment. It highlighted the branching decisions that lead to classifications of default or non-default. The algorithm’s performance, validated by accuracy and cross-validation scores, demonstrates its robustness and reliability even amidst economic fluctuations.

Implications for Credit Risk Management: The interpretations derived from our models suggest that more emphasis should be placed on the debt-to-income ratio when evaluating creditworthiness. Additionally, it is apparent that incorporating a variety of factors into the risk assessment process can yield a more holistic view of an applicant's financial health.

Model Selection and Credit Scoring: In selecting a model that best addresses our research questions, we prioritize not only accuracy but also interpretability. The Random Forest model, while highly accurate, offers less interpretability than the Decision Tree model, which provides a clear and concise representation of the decision-making rules. This is particularly valuable for creating an understandable and actionable credit scorecard.

Conclusion: Our analysis has provided answers to our research questions, uncovering significant determinants of credit risk, revealing patterns within the data, and validating the effectiveness of the decision tree algorithm during economic instability. These insights will serve as a foundation for developing a robust credit scoring system that can withstand the complexities of economic uncertainty.

Limitations of the research:

While the research provides substantial insights into credit risk assessment, there are several limitations that must be acknowledged:

Model Bias and Generalizability:

The predictive models developed may have inherent biases due to the nature of the training data. If the dataset is not representative of the broader population or if it reflects historical biases, the models' predictions may not generalize well to other groups or scenarios.

Feature Selection and Data Availability:

The accuracy of the models is highly dependent on the chosen features. There might be other variables not included in the dataset that could significantly affect credit risk assessment. Additionally, any missing or inaccurate data can lead to less reliable predictions.

Interpretability vs. Accuracy Trade-Off:

There is a trade-off between model complexity and interpretability. More complex models like Random Forest provide better accuracy but at the cost of interpretability, which is crucial for understanding the drivers of credit risk.

Risk of Overfitting:

There is a risk that the models may overfit to the training data, capturing noise as a signal, which could lead to poor performance on unseen data.

External Factors:

Factors such as changes in legislation, global economic crises, or pandemics are external variables that can affect an individual's credit risk and are not accounted for in the models.

In conclusion, these limitations underscore the need for cautious interpretation of the model results and suggest areas for future research. It is important for subsequent studies to address these limitations by incorporating a wider range of data, considering the evolving economic landscape, and balancing the accuracy of predictions with the ethical implications of automated decision-making in credit risk assessment.

Remarks on Continuity and Areas for Improvement

Despite the limitations, this study establishes a groundwork for an in-depth analysis of credit risk assessment using machine learning. Here are some recommendations for future research to enhance the robustness and applicability of the findings:

Expanding Data Collection:

Future research should aim to collect a more diverse set of data points, including more granular information about borrowers' financial behavior, economic variables, and demographic factors that may influence creditworthiness.

Dynamic Model Updating:

Developing a mechanism for continuous learning where models are regularly updated with new data can help maintain their accuracy over time and adapt to changing economic conditions.

Incorporating Alternative Data:

Exploring non-traditional data sources, such as utility payments, rental history, or even social media activity, could uncover additional predictors of credit risk that are not captured by traditional financial metrics.

Enhancing Model Interpretability:

Investing in explainable AI techniques will make it easier to interpret complex models, which is critical for gaining stakeholder trust and for the models to be actionable.

Real-time Analysis Capabilities:

Developing systems that can analyze credit risk in real-time, taking into account the most recent data, can provide more accurate and timely assessments.

Longitudinal Studies:

Long-term studies that track the performance of credit risk models over time can provide insights into their durability and the changing patterns of creditworthiness.

In conclusion, continued research in these areas promises not only to refine the predictive power of credit risk assessment models but also to ensure their ethical application and relevance in the face of rapidly evolving financial landscapes. By addressing these areas for improvement, future research can build more robust, fair, and transparent models that better serve the needs of both lenders and borrowers.

Conclusion

In this capstone project, I embarked on an analytical expedition to develop a predictive model for credit risk assessment—a model that needed to be precise, interpretable, and practically applicable within the financial lending domain. Through an intensive exploration of over 32,000 loan transactions, I have successfully architected a Random Forest model that not only showcases a high degree of accuracy in predicting defaults but also integrates seamlessly with user-friendly credit risk scorecards.

My research was guided by three pivotal questions that steered my investigation towards understanding the factors influencing credit defaults, constructing a transparent scorecard, and verifying the model's reliability. The Random Forest model emerged as a standout, demonstrating its proficiency through high accuracy and precision rates. It flagged essential predictors such as the loan-to-income ratio and credit history length—factors that are vital to risk evaluations and prudent lending.

Selecting and endorsing the final model was a balanced act of considering statistical validation and the practical context of its application. The interpretability of the Decision Tree model was particularly notable for its potential to convert complex statistical findings into understandable decision-making rules.

The conclusion of this project marks the beginning of a continuous journey. The implementation of the model is just the first step, with ongoing monitoring and updates being crucial to its sustained success in light of economic shifts and market dynamics.

The insights and methodologies applied here pave the way for future inquiries and advancements. This project illustrates the transformative potential of data analytics in the financial sector, signifying a stride towards a more stable and data-informed lending environment.

As I conclude, I see this project not just as a successful academic endeavor but also as a cornerstone for future innovation—a testament to the transformative power of data science in financial decision-making and a beacon for risk mitigation in the evolving landscape of financial services.

References

Si Shi1 • Rita Tse1,2 • Wuman Luo1 • Stefano D'Addona3 • Giovanni Pau4,5 (2022). Machine learning-driven credit risk: a systemic review.

https://www.researchgate.net/publication/362051812_Machine_learning-driven_credit_risk_a_systemic_review

Paritosh Navinchandra Jha 1,* and Marco Cucculelli 2 (2021). A New Model Averaging Approach in Predicting Credit Risk Default. Risks | Free Full-Text | A New Model Averaging Approach in Predicting Credit Risk Default (mdpi.com)

David Mhlanga (2021). Financial Inclusion in Emerging Economies: The Application of Machine Learning and Artificial Intelligence in Credit Risk Assessment. Financial Inclusion in Emerging Economies: The Application of Machine Learning and Artificial Intelligence in Credit Risk Assessment – DOAJ

Julapa Jagtiani, Catharine Lemieux (2017). Fintech Lending: Financial Inclusion, Risk Pricing, and Alternative Information. Fintech Lending: Financial Inclusion, Risk Pricing, and Alternative Information by Julapa Jagtiani, Catharine Lemieux :: SSRN

Sarah Kornfeld (2020). Predicting Default Probability in Credit Risk using Machine Learning Algorithms. FULLTEXT01.pdf (diva-portal.se)