

Credit Risk Default Prediction

Harshit Katheria

24/08/2023

1. Abstract:

Credit default remains a significant concern within the financial landscape, impacting businesses and individuals alike. Timely prediction of credit default holds the potential to mitigate financial risks and optimize lending decisions. This report is centered around developing a machine learning-driven framework for credit default prediction based on relevant financial attributes. The objective is to construct an accurate and efficient system capable of assisting financial institutions in evaluating the likelihood of credit default. By analysing a range of financial parameters, this study aims to forecast whether a credit applicant is prone to defaulting, employing a suite of machine learning classification models including Logistic Regression, Decision Trees, Random Forest, and more.

2. Problem Statement:

Since the financial crisis, companies have come to understand the significance of risk management with the latest technology. To date, machine learning algorithms have been applied in financial research and financial service industry. Essentially the use of credit-scoring algorithms helps financial service providers estimate the creditworthiness of borrowers to reduce the labour costs and constantly maintains sustainable development of the financial world. As such, a critical problem emerges: How can the financial industry effectively harness the power of machine learning algorithms for credit scoring while addressing ethical and practical challenges associated with algorithmic bias, ensuring a reliable eco-system for the consumers and its stakeholders.

3. Market/Customer/Business Need Assessment:

1. **Market Analysis:** Understanding the market landscape is crucial. This involves identifying the target audience for credit default prediction system, such as financial institutions, lenders, or credit agencies. Analyzing the competitive landscape and existing solutions in the market helps identify gaps and opportunities for differentiation.
2. **Business Objectives:** Aligning the system with business objectives is important. The assessment should consider how the credit default system supports to company's goals, such as minimizing loan defaults, improving portfolio quality, enhancing operational efficiency and increasing profitability.

3. **Value Proposition:** Define the unique value of credit default prediction system offers. This could include improved loan approval process, reduced default rates, better risk management and enhanced customer satisfaction.

Meeting the market, customer, and business needs for a credit risk default prediction system involves creating and deploying reliable and precise models that seamlessly integrate with financial operations, offer scalability, generate actionable predictions, and align with regulatory and ethical guidelines. Effective and efficient achievement of these needs requires close collaboration among financial institutions, technology experts, regulatory authorities, and relevant stakeholders.

4. *Target Specification and Characterization:*

1. **Accuracy and Precision:** The model's ability to accurately predict credit default events and distinguish between various levels of risk.
2. **Predictive Factors:** Identifying crucial features such as monthly income, number of dependents, open credit lines, debt ratios and other financial indicators that contribute to predicting credit default.
3. **Prediction Reliability:** Ensuring a dependable level of accuracy based on different credit default scenarios and financial conditions.
4. **Model Architecture:** The model's architecture should be designed to effectively handle the complexity and variability of various kinds of credit default scenarios.
5. **Robustness:** The model should exhibit robust performance across different datasets
6. **Prediction Accessibility:** Evaluating the accessibility of the prediction model, considering factors like implementation costs, processing time, and availability across various financial sectors.

5. *External search:*

As we all know that the dataset plays a crucial role in Machine Learning. Availability of proper dataset is not very common. To help us out Kaggle gave us a large amount of data set. The Dataset for credit default prediction (GiveMeSomeCredit) is available on Kaggle. The dataset is designed in such a way that is easy to understand. Having various columns like monthly income, number of dependents, number of open credit lines and loans, debt ratio & age. Target variable which is binary class containing the result of if a person will default on his payments or not (positive, negative).

<https://www.kaggle.com/c/GiveMeSomeCredit/discussion?sort=votes-patients-and-air-pollution-anew-link>

6. Bench marking alternative products:

In the realm of credit risk default prediction, while there may not be specific competing products, financial institutions have historically relied on traditional credit assessment methods. Our machine learning-based model offers distinct advantages:

1. **Cost-Efficiency:** Our solution is more economical compared to conventional assessment methods used by financial institutions.
2. **Ease of use:** The model is designed for user ease, enhancing decision-making confidence.
3. **Accuracy:** As this device is completely based on AI and ML approaches its accuracy is always high.
4. **Time Efficiency:** Accessible through web applications, it not only serves the general public but also saves time for decision-makers.
5. **Wide Availability:** By being accessible on various platforms, including E-commerce sites, it ensures availability and convenience for all.

7. Applicable Regulations:

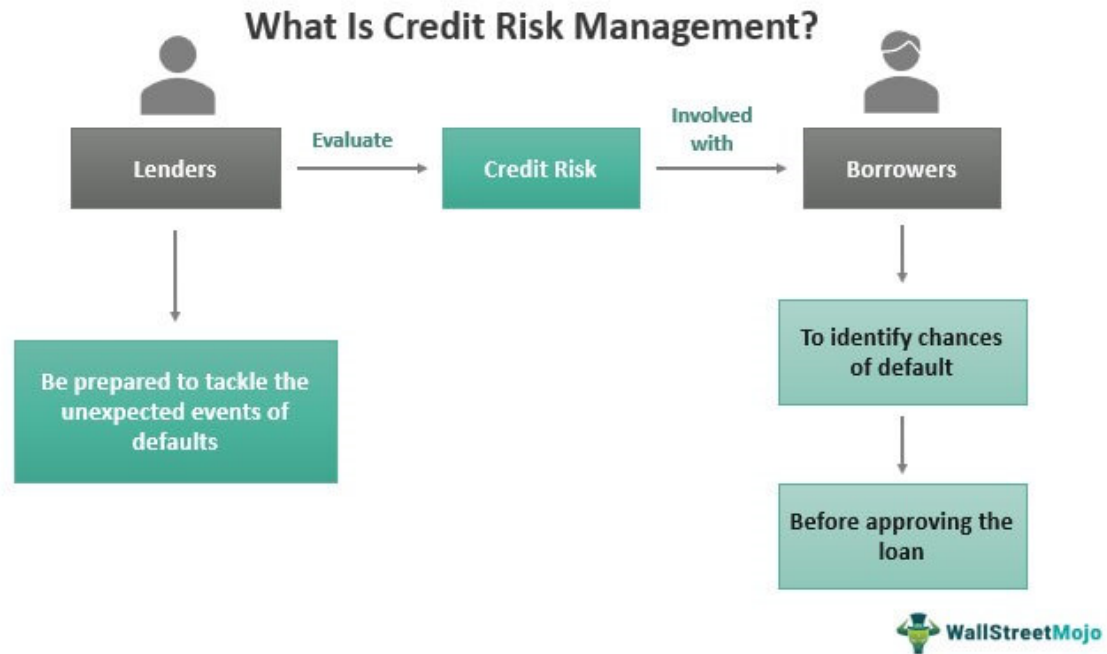
The deployment of credit risk prediction models in India is subject to regulatory guidelines aimed at maintaining transparency, fairness, and data privacy. Financial institutions and lending entities must comply with these regulations to ensure responsible and ethical use of such models. Some of the relevant regulations include:

1. Reserve Bank of India (RBI) Guidelines: The RBI provides guidelines and directives to financial institutions regarding credit risk management, ensuring that lending practices are prudent and risk-appropriate. The Indian Medical Devices (Marketing Authorization) Regulations, 2017
2. Credit Information Companies (Regulation) Act, 2005: This act governs the operations of credit information companies in India, which collect and maintain credit-related information about individuals and businesses. It ensures that the data is accurate, updated, and used only for authorized purposes.
3. Prevention of Money Laundering Act, 2002: Financial institutions are required to follow anti-money laundering and know-your-customer (KYC) norms to prevent the misuse of financial services for illicit purposes.

8. Applicable Constraints:

1. **Data Requirement:** Since whole device is completely based on Data , so proper availability of data is required and it must be accurate, which become a great challenge for Data scientist.
2. **Technical Complexity:** Machine Learning Algorithms are complex and require extensive data and computational resources to develop and deploy effectively. The development and deployment of the device may also require specialized expertise in machine learning, data science, and finance.
3. **Regulatory Compliance:** The credit risk prediction system must adhere to regulatory standards and guidelines set by financial authorities. Compliance with regulations such as data privacy laws and financial regulations is crucial to ensure the system's legality and trustworthiness.
4. **Cost Considerations:** The development, deployment, and maintenance of the credit risk prediction system can involve significant costs. This includes expenses related to software development, data collection, model training, and ongoing monitoring. Additionally, training financial professionals to effectively use the system can also add to the costs
5. **Trust and Validation:** The credit risk prediction system's accuracy and reliability play a crucial role in gaining trust from financial institutions and customers. Rigorous validation and testing are essential to ensure that the system provides accurate predictions that align with expert assessments.
6. **Ethical Considerations:** The use of machine learning for credit risk assessment raises ethical concerns related to bias, fairness, and transparency. It's important to address potential biases in the data and algorithms to ensure that the system does not discriminate against certain groups of individuals.

9. BUSINESS MODEL:



In the initial phase, our strategy involves building the Credit Risk Default Prediction model into a user-friendly web application. This approach ensures widespread accessibility and user engagement. To generate revenue, we'll leverage online advertising on the platform. As users interact with the web app, not only will the dataset grow, enhancing accuracy, but also the product's presence will expand through word-of-mouth. This phase aims to establish the model's credibility and demonstrate its ability to provide accurate predictions and valuable suggestions to users.

Expanding into the second phase, our focus will shift towards developing a mobile app version of the Credit Risk Default Prediction model. This mobile app will provide users with the convenience of accessing credit risk predictions on-the-go. To monetize the app, we'll introduce a subscription model offering advanced features, such as detailed credit reports and personalized financial recommendations. This phase aims to enhance user engagement and establish a recurring revenue stream.

The third phase involves integrating the Credit Risk Default Prediction model into financial institutions, such as banks and lending companies. This integration will empower institutions to assess credit risk more accurately and make informed lending decisions. To generate revenue, we'll establish collaborations using a B2B partnership model. Financial institutions subscribing to our service will pay a licensing fee for accessing our predictive analytics. This phase targets a broader impact on the financial sector and solidifies the model's credibility.

10. Concept Generation:

1. **Identifying the problem statement:** The initial step is to clearly understand and define the problem we aim to solve. In this case, the key issue addressed by our model is

predicting credit default risk based on various financial factors such as income, debt ratios, and credit history.

2. **Define the goals and objectives:** Setting clear goals and objectives is essential for the successful development of our credit risk prediction model. Our objectives include creating a model that accurately predicts credit default risk, helps lenders make informed decisions, and reduces financial losses.
3. **Determining feasibility and impact:** We evaluate the feasibility of our solution by considering factors like data availability, technical resources, legal considerations, and potential challenges. Simultaneously, we gauge the impact of our model by examining its potential benefits, such as improved lending decisions, reduced defaults, and enhanced financial stability
4. **Choose the appropriate machine learning algorithms and techniques:** Choosing the appropriate machine learning algorithm is also very crucial because this will determine our device's accuracy. Various types of learning will be used like supervised and unsupervised learning, ensemble methods. The choice of algorithms will be based on the characteristics of the credit data and the desired outcome.
5. **Define the data requirements:** As data quality is paramount, we outline the specific data needs for our model. This includes variables like credit history, income, loan amount, and repayment behaviour. We may need to source data from credit bureaus, financial institutions, and economic indicators. This may involve collecting and labelling new data, accessing existing datasets, or partnering with healthcare organizations to obtain the necessary data.
6. **Design the system architecture:** The next step involves selection of all those hardware and software components over which our device will work. This may involve cloud-based or on-premise solution, selection of programming language and frameworks. Scalability and efficiency will also matter.
7. **Develop a prototype:** In the final stage, we build a prototype of our credit risk prediction system. This involves designing a user-friendly interface for inputting financial data, integrating data pipelines, and conducting rigorous testing. The prototype's accuracy, speed, and overall performance are validated before moving to deployment

11. Concept Development:

Our Credit Risk Default Prediction model utilizes advanced machine learning algorithms to analyse financial data and predict the likelihood of credit default. By incorporating various financial factors like income, debt ratios, and credit history, the model identifies intricate patterns and relationships that might elude human analysts. This pattern recognition is often complex and beyond the scope of traditional analysis.

The system's AI-driven approach ensures swift and precise risk assessment, enhancing the efficiency of lending processes. Moreover, it contributes to reducing financial losses by enabling lenders to make well-informed decisions. The development of this model requires a collaborative

effort from diverse experts, including data scientists, financial professionals, data analysts, software developers, and project managers.

The development process involves careful selection of appropriate machine learning algorithms, frameworks, and software tools. Data must be accurately labeled and pre-processed to ensure reliable predictions. Best practices in software development and deployment are adhered to, ensuring the model's accuracy and reliability. As this model significantly impacts financial decisions, it undergoes rigorous validation, adheres to regulatory standards, and aligns with ethical considerations to safeguard consumer interests and regulatory compliance

12. Final Product Prototype:

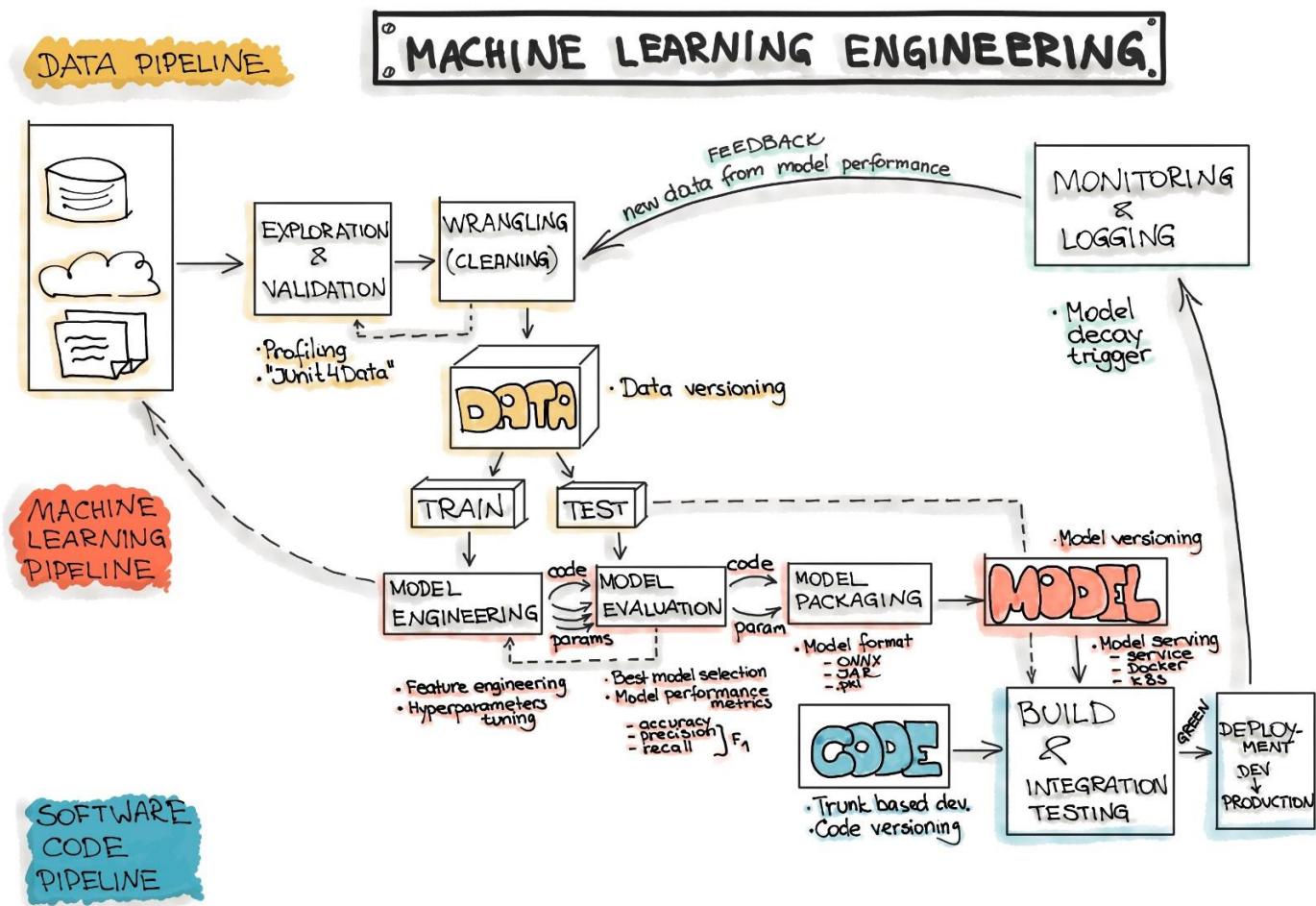
1. **User interface:** The user interface is meticulously designed for a seamless and user-friendly experience, ensuring easy navigation and accessibility for users.
2. **Data collection and preprocessing:** The system gathers relevant financial and credit-related data from various sources, preprocesses the data by handling missing values and outliers, and prepares it for analysis.
3. **Machine Learning Algorithm:** Employing a range of advanced classification algorithms, the system delves into the data to identify intricate patterns and relationships that contribute to accurate credit default prediction.
4. **Prediction report:** The system generates comprehensive credit risk prediction reports, offering insights into the likelihood of default for individuals or entities along with factors influencing the prediction
5. **Performance metrics:** Key performance metrics such as accuracy, precision, recall, and F1-score are included to quantify the effectiveness of the machine learning model's predictions.
6. **Regulatory compliance:** The prototype adheres to relevant regulatory guidelines and data protection laws, ensuring the security and privacy of sensitive financial and personal information.
7. **Scalability:** To handle the big amount of data its scalability would be handled so that it never fails to train over big amount of data.
8. **Cloud based or on -premise solution:** Depending upon the demand of organization the device could be designed as a cloud-based or on-premise solution.

13. Product Details:

1. Working:

1. **Data collection:** The initial step involves gathering a comprehensive and diverse dataset containing various factors associated with credit risk. This dataset encompasses historical credit data, financial attributes, borrower information, repayment history, and other relevant variables.

2. **Data Preprocessing:** The collected data undergoes preprocessing to ensure its quality and usability for machine learning algorithms. This step includes eliminating duplicate or irrelevant information, addressing missing values, and standardizing data formats.
3. **Feature Engineering:** Extracting meaningful features from the preprocessed data is essential for accurate credit risk assessment. Statistical methods are employed to identify critical attributes that contribute significantly to predicting credit default.
4. **Model Selection:** Given that credit default prediction involves classification, appropriate classification algorithms are chosen. Common choices include decision trees, Random Forest, logistic regression, Support Vector Machines (SVM), and advanced techniques like XGBoost or LightGBM.
5. **Model Training:** During this phase, the selected model is trained using the preprocessed data. The model learns intricate patterns, relationships among attributes, and relevant factors influencing credit default predictions.
6. **Model Evaluation:** This is the most important phase where our model is tested over the unseen data and its prediction is evaluated. Based on which we determine how much accuracy our model provide.
7. **Deployment:** Once the model achieves the desired accuracy, it's ready for deployment. The credit risk prediction model can be integrated into a software application, web platform, or existing financial systems used by banks, lending institutions, or credit assessment agencies.



2.Data Sources:

- Credit Data Providers:** These providers offer historical credit data including borrower information, payment history, credit limits, and outstanding balances. This data forms the foundation for training the credit risk prediction mode.
- Financial Institutions' Records:** Data from banks, lending companies, and credit card companies contain valuable information about borrowers' financial behaviors, loan applications, and repayment patterns.
- Credit Bureaus:** Credit bureaus compile credit reports containing credit scores, account histories, and credit inquiries. This data aids in assessing borrowers' creditworthiness
- Industry Specific Data:** For specialized lending sectors, industry-specific data such as real estate market trends, automotive sales, or commodity prices may be relevant for credit risk assessment.

3. Algorithms, frameworks and software needs:

1. **Algorithm:** The credit default prediction model employs a variety of machine learning algorithms, primarily focusing on supervised classification methods like logistic regression, decision trees, random forest, gradient boosting, XGBoost and LightGBM.
2. **Frameworks:** The development of the credit risk prediction model involves the utilization of python, scikit-learn, Flask or Django.
3. **Software:** To handle the large dataset we need cloud storage or cloud platforms like AWS, GCP, Microsoft Azure to run heavy models. IDE's like Pycharm, Jupyter Notebook, Google Colab and others. To perform the version control we use Git and Github so that we can collaborate with others developers. For developing, shipping and running application we can use Docker and Kubernetes.
4. **Visualization tools:** Matplotlib and seaborn can be used to data visualization and data exploration.

4. Team Required to Develop

1. **Data Scientist and ML engineers:** These experts specialize in machine learning algorithms, data preprocessing, feature engineering, and model training. They are responsible for creating the predictive model using historical credit data.
2. **Domain Experts:** Professionals from the financial sector, such as credit risk analysts and financial advisors, provide valuable domain knowledge and insights to guide the model's development and ensure its alignment with industry standards.
3. **Data analyst and statisticians:** They play a vital role in examining and preparing the data for analysis. They identify patterns, trends, and correlations within the credit-related data, which contribute to the model's accuracy.
4. **Software developer:** These individuals are responsible for implementing the predictive model into software applications, whether they are web-based interfaces or integrated into existing systems. They ensure seamless integration and user-friendly interfaces.
5. **Project Managers:** Effective project management is crucial for coordinating the efforts of the various team members, setting milestones, managing timelines, and ensuring that the project stays on track and within budget.
6. **Quality Assurance Specialist:** QA specialists rigorously test the model to identify any bugs, inconsistencies, or inaccuracies. They ensure that the model's predictions align with expectations and deliver reliable results.

5. What does it Cost?

Developing and implementing a credit default prediction model involves several cost-driving factors that contribute to the overall expenses of the project:

1. **Data Collection and Preprocessing:** Acquiring accurate and comprehensive credit-related data, including financial records and historical payment behavior, might require partnerships with data providers or access to credit bureau databases, which can incur costs.
2. **Hard ware and software infrastructure:** Acquiring accurate and comprehensive credit-related data, including financial records and historical payment behavior, might require partnerships with data providers or access to credit bureau databases, which can incur costs.
3. **Regulatory Compliance:** Ensuring compliance with regulatory standards, such as data protection regulations and industry-specific guidelines, often involves legal consultations and auditing processes. The costs associated with achieving and maintaining compliance can be significant.
4. **Expertise and resources:** Building and maintaining a proficient team of data scientists, domain experts, software developers, and project managers requires a financial commitment for salaries, benefits, training, and professional development.
5. **Testing and Validation:** Rigorous testing and validation of the credit default prediction model demand resources for setting up controlled testing environments, conducting real-world simulations, and verifying the model's accuracy.
6. **Deployment and Integration:** Integrating the model into existing systems, whether they are web-based applications or software solutions, requires development, testing, and potentially third-party integration tools.
7. **Ethical Considerations:** Incorporating ethical considerations, especially in finance-related applications, might lead to additional costs associated with ensuring fair and unbiased decision-making.
8. **Maintenance and Update:** Continuously updating and maintaining the model's accuracy, performance, and compliance with changing regulations can lead to ongoing costs.
9. **Marketing and Outreach:** To effectively introduce the credit risk prediction system to the market, marketing efforts, including promotional activities and campaigns, will incur expenses.
10. **Customer Support:** Providing efficient customer support services, addressing inquiries, and resolving issues can require dedicated personnel and resources.
11. **Training and Education:** Educating users, stakeholders, and employees about the system's features, benefits, and proper usage might involve costs related to training materials, workshops, and resources.

14. Code Implementation:

1. Importing Libraries and Loading Dataset:

```
import jovian
import pandas as pd
import numpy as np
import warnings
warnings.filterwarnings("ignore")
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
```

[] ⓘ Python

...

▶ ▾ train_df.info()

[4] Python

... <class 'pandas.core.frame.DataFrame'>
RangeIndex: 150000 entries, 0 to 149999
Data columns (total 12 columns):

#	Column	Non-Null Count	Dtype
0	Unnamed: 0	150000 non-null	int64
1	SeriousDlqin2yrs	150000 non-null	int64
2	RevolvingUtilizationOfUnsecuredLines	150000 non-null	float64
3	age	150000 non-null	int64
4	NumberOfTime30-59DaysPastDueNotWorse	150000 non-null	int64
5	DebtRatio	150000 non-null	float64
6	MonthlyIncome	120269 non-null	float64
7	NumberOfOpenCreditLinesAndLoans	150000 non-null	int64
8	NumberOfTimes90DaysLate	150000 non-null	int64
9	NumberRealEstateLoansOrLines	150000 non-null	int64
10	NumberOfTime60-89DaysPastDueNotWorse	150000 non-null	int64
11	NumberOfDependents	146076 non-null	float64

dtypes: float64(4), int64(8)

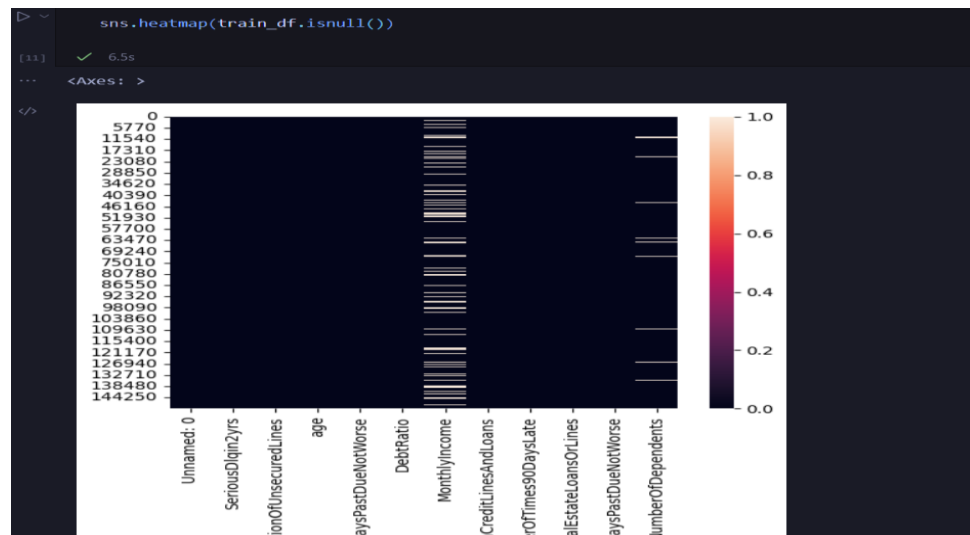
2. Data Preprocessing:

```
Data Preprocessing

train_df.isnull().sum()

[4]

...   Unnamed: 0      0
      SeriousDlqin2yrs      0
      RevolvingUtilizationOfUnsecuredLines      0
      age      0
      NumberOfTime30-59DaysPastDueNotWorse      0
      DebtRatio      0
      MonthlyIncome      29731
      NumberOfOpenCreditLinesAndLoans      0
      NumberOfTimes90DaysLate      0
      NumberRealEstateLoansOrLines      0
      NumberOfTime60-89DaysPastDueNotWorse      0
      NumberOfDependents      3924
      dtype: int64
```



```
from sklearn.impute import SimpleImputer
cols_with_missing_values = ['MonthlyIncome', 'NumberOfDependents']

✓ 1.6s

imputer = SimpleImputer(strategy='mean')

✓ 0.0s

train_df[cols_with_missing_values] = imputer.fit_transform(train_df[cols_with_missing_values])

✓ 0.0s

train_df.isnull().sum()

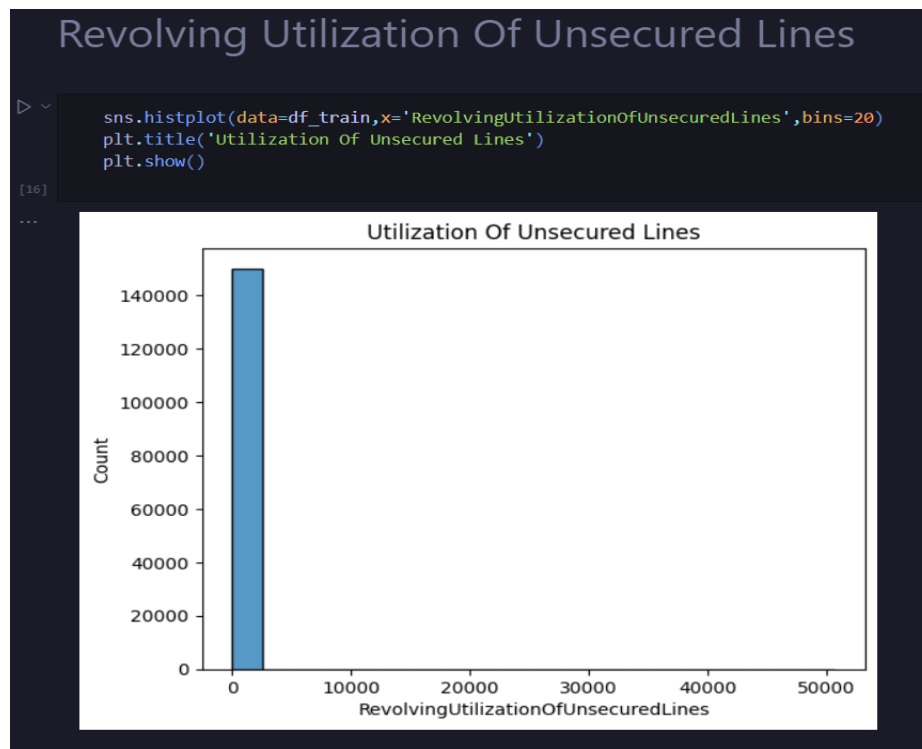
✓ 0.0s

Unnamed: 0      0
SeriousDlqin2yrs      0
RevolvingUtilizationOfUnsecuredLines      0
age      0
NumberOfTime30-59DaysPastDueNotWorse      0
DebtRatio      0
MonthlyIncome      0
NumberOfOpenCreditLinesAndLoans      0
NumberOfTimes90DaysLate      0
NumberRealEstateLoansOrLines      0
NumberOfTime60-89DaysPastDueNotWorse      0
NumberOfDependents      0
dtype: int64
```

3. Final Dataset:

Final DataSet						
train_df						
[29]	✓	0.0s				
...						
0	1	SeriousDlqin2yrs	RevolvingUtilizationOfUnsecuredLines	age	\	
1	2		0.766127	45		
2	3		0.957151	40		
3	4		0.658180	38		
4	5		0.233810	30		
...	0.907239	49		
149995	149996		0.040674	74		
149996	149997		0.299745	44		
149997	149998		0.246044	58		
149998	149999		0.000000	30		
149999	150000		0.850283	64		
		NumberOfTime30-59DaysPastDueNotWorse	DebtRatio	MonthlyIncome	\	
0		2	0.802982	9120.000000		
1		0	0.121876	2600.000000		
2		1	0.085113	3042.000000		
3		0	0.036050	3300.000000		
4		1	0.024926	63588.000000		
...		
149995		0	0.225131	2100.000000		
149996		0	0.716562	5584.000000		
149997		0	3870.000000	6670.221237		
149998		0	0.000000	5716.000000		
149999		0	0.249908	8158.000000		
...						
149998		0.0				
149999		0.0				

4. Exploratory Data Analysis:

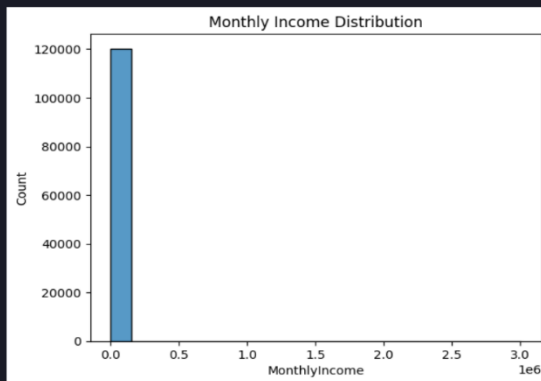


Monthly Income

```
sns.histplot(data=df_train,x='MonthlyIncome',bins=20)
plt.title('Monthly Income Distribution')
plt.show()
```

[25]

Python

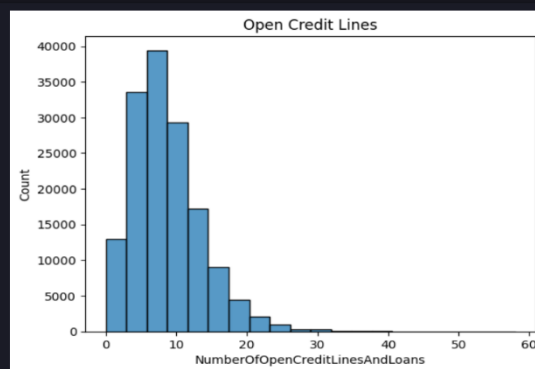


Number Of Open Credit Lines And Loans

```
sns.histplot(data=df_train,x='NumberOfOpenCreditLinesAndLoans',bins=20)
plt.title('Open Credit Lines')
plt.show()
```

[26]

Python

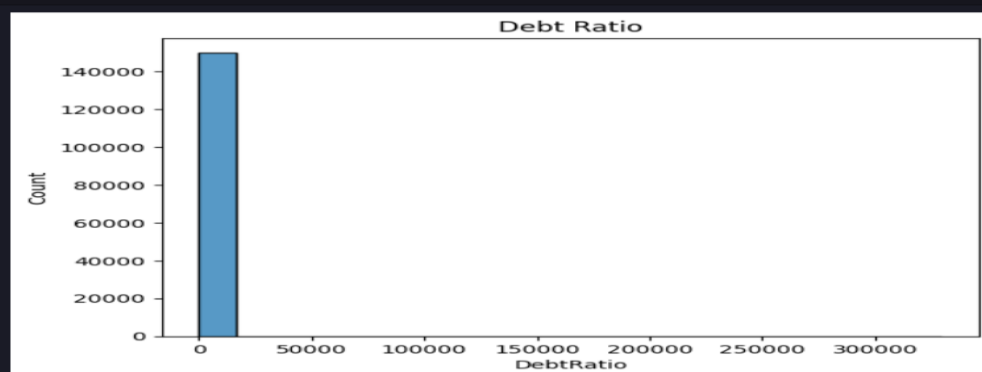


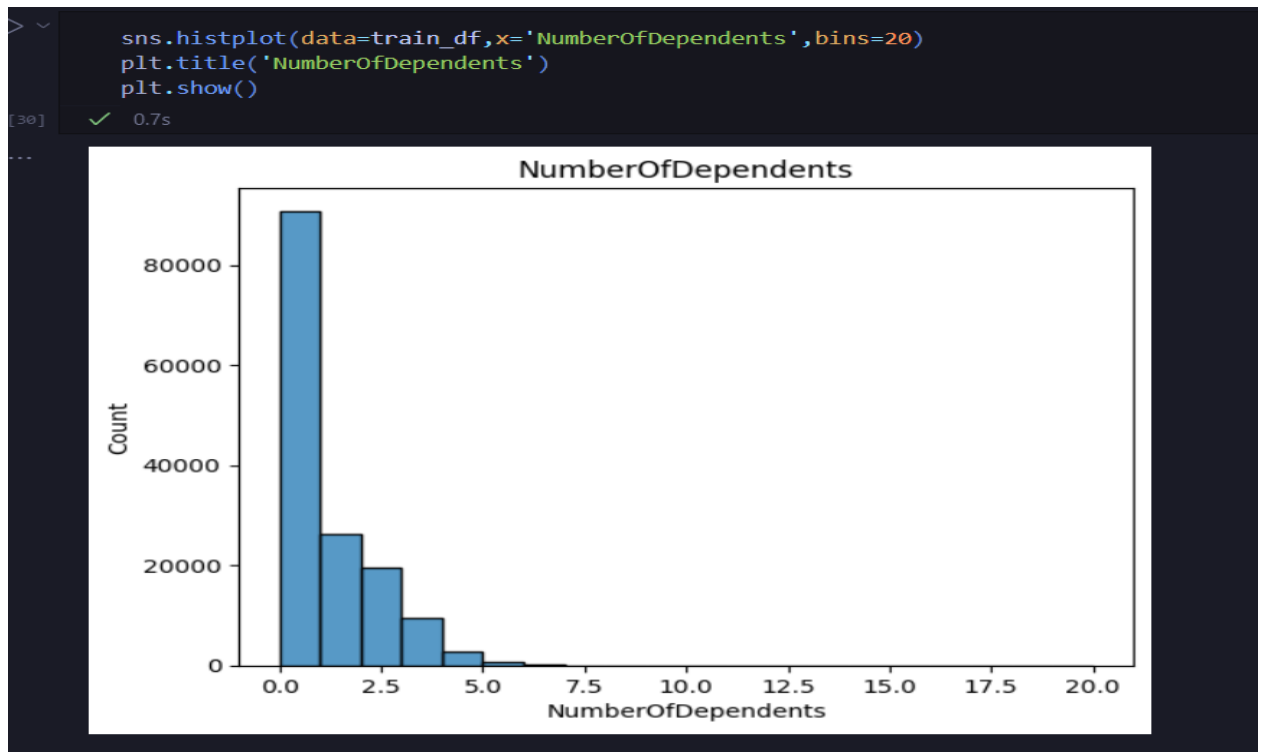
DebtRatio

```
sns.histplot(data=df_train,x='DebtRatio',bins=20)
plt.title('Debt Ratio')
plt.show()
```

[27]

...





Summary of EDA Findings:

Unusual Data Patterns and Outliers: The dataset exhibits several anomalies and outliers that deviate from expected financial behaviors. These anomalies include unrealistic values in features like *RevolvingUtilizationOfUnsecuredLines*, *DebtRatio*, and *MonthlyIncome*, which are indicative of potential data entry errors, extreme financial situations, or missing data. Such instances present challenges for model training and require special attention.

Complex Relationships and Missing Data: Key features like *DebtRatio* and *MonthlyIncome* show complex relationships and missing data points. High *DebtRatio* values sometimes coincide with missing *MonthlyIncome*, and the patterns are consistent in both the training and test datasets. Establishing a clear relationship between these variables is difficult due to the missing data and inconsistent trends, making it challenging to filter out anomalies.

Credit Line and Loan Dynamics: The feature indicating the number of open credit lines and loans presents challenges in determining the distinction between sources of income and debts. While some instances involve implausible values or high numbers of dependents, filtering the data might not be appropriate, as similar patterns are observed in the test dataset. Handling such cases is important to avoid potential defaults, and training the model to learn from these anomalies may lead to better predictions.

5. Splitting the Dataset into training and validation sets:

```
[37] from sklearn.model_selection import train_test_split

X_train_data = df_train.drop('SeriousDlqin2yrs',axis=1)
y_train_data = df_train['SeriousDlqin2yrs']
X_test_data = df_test.drop(['SeriousDlqin2yrs'], axis=1)
y_test_data = df_test['SeriousDlqin2yrs']

[38]

X_train, X_val, y_train, y_val = train_test_split(X_train_data, y_train_data, test_size=0.33, random_state=42)

[39]
```

6. Training of Machine Learning Algorithms:

Decision Trees

```
▷ ▾ #DecisionTrees
from sklearn.tree import DecisionTreeClassifier
model = DecisionTreeClassifier(random_state=42)
model.fit(X_train,y_train)
train_preds = model.predict(X_train)
val_preds = model.predict(X_val)
train_accuracy = accuracy_score(y_train,train_preds)
val_accuracy = accuracy_score(y_val,val_preds)
f1 = f1_score(y_val,val_preds)
recall = recall_score(y_val,val_preds)
precision = precision_score(y_val,val_preds)

print("Training Accuracy:", train_accuracy)
print("Validation Accuracy:", val_accuracy)]
print("F1 Score:", f1)
print("Precision:", precision)
print("Recall:", recall)

[63] ✓ 4.5s

... Training Accuracy: 1.0
Validation Accuracy: 0.9004646464646464
F1 Score: 0.27104601272377576
Precision: 0.2612664004563605
Recall: 0.2815862280971411
```

RandomForest

```
#RandomForest
model = RandomForestClassifier(random_state=42)
model.fit(X_train,y_train)
train_preds = model.predict(X_train)
val_preds = model.predict(X_val)
train_accuracy = accuracy_score(y_train,train_preds)
val_accuracy = accuracy_score(y_val,val_preds)
f1 = f1_score(y_val,val_preds)
recall = recall_score(y_val,val_preds)
precision = precision_score(y_val,val_preds)

print("Training Accuracy:", train_accuracy)
print("Validation Accuracy:", val_accuracy)
print("F1 Score:", f1)
print("Precision:", precision)
print("Recall:", recall)
```

[66] ✓ 1m 48.9s

```
... Training Accuracy: 0.9999701492537313
Validation Accuracy: 0.9372121212121212
F1 Score: 0.28055555555555556
Precision: 0.5679475164011246
Recall: 0.18628957885029204
```

LightGBM

+ Code

+ Markdown

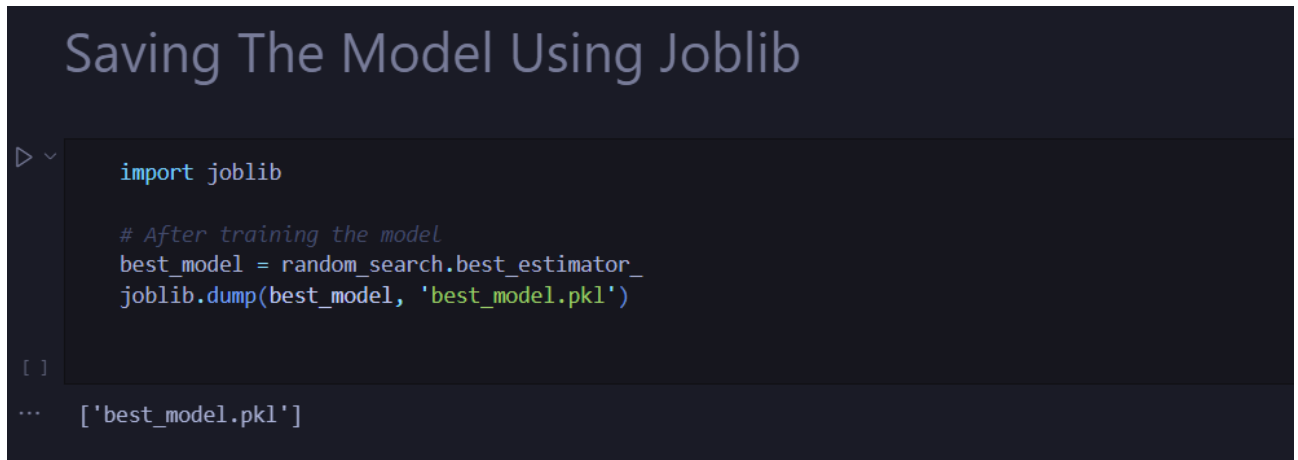
```
#LightGBM
model = lgb.LGBMClassifier()
model.fit(X_train,y_train)
train_preds = model.predict(X_train)
val_preds = model.predict(X_val)
train_accuracy = accuracy_score(y_train,train_preds)
val_accuracy = accuracy_score(y_val,val_preds)
f1 = f1_score(y_val,val_preds)
recall = recall_score(y_val,val_preds)
precision = precision_score(y_val,val_preds)

print("Training Accuracy:", train_accuracy)
print("Validation Accuracy:", val_accuracy)
print("F1 Score:", f1)
print("Precision:", precision)
print("Recall:", recall)
```

[]

```
... [LightGBM] [Info] Number of positive: 6773, number of negative: 93727
[LightGBM] [Warning] Auto-choosing row-wise multi-threading, the overhead of testing was 0.006241 seconds.
You can set `force_row_wise=true` to remove the overhead.
And if memory is not enough, you can set `force_col_wise=true`.
[LightGBM] [Info] Total Bins 1227
[LightGBM] [Info] Number of data points in the train set: 100500, number of used features: 11
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.067393 -> initscore=-2.627442
[LightGBM] [Info] Start training from score -2.627442
Training Accuracy: 0.9417910447761194
Validation Accuracy: 0.9377777777777778
F1 Score: 0.29422548120989916
Precision: 0.5778577857785778
Recall: 0.19735628650476483
```

Saving the model locally:



```
import joblib

# After training the model
best_model = random_search.best_estimator_
joblib.dump(best_model, 'best_model.pkl')

[ ]

... ['best_model.pkl']
```

15. Future Updation:

In our credit default prediction system, our focus has primarily been on analyzing various financial and credit-related factors. However, there are several avenues for enhancing the system's accuracy and capabilities. To further improve prediction accuracy, we can integrate additional data sources such as social media activity, transaction history, and economic indicators. This can provide a comprehensive view of an individual's financial behaviour.

Developing intuitive and user-friendly interfaces for both professionals and consumers will encourage wider adoption of the system's insights.

16. Conclusion:

In conclusion, our credit default prediction model stands as a powerful tool for financial institutions and businesses to assess and manage credit risk effectively. By leveraging advanced machine learning algorithms and data-driven insights, this model can aid in making informed lending decisions, minimizing default risks, and optimizing financial strategies. However, it's crucial to remember that while this model can offer valuable predictions, it should be complemented by human expertise and industry knowledge. Striking the right balance between automation and human judgment is key to maximizing the model's utility.

As we move forward, the continuous refinement and enhancement of our credit default prediction system will be vital. The dynamic nature of the financial landscape and the ever-evolving economic conditions require us to adapt and innovate. Regular updates, incorporation of new data sources,

and collaboration with financial institutions will ensure that our model remains relevant and effective in helping businesses navigate the complex world of credit risk

[Project Link](#)