**Credit Risk Default Prediction**

By

Pritam Rajak

Harshit Katheria

Pranav Sunil Jadhav

Date: 24th October 2023



**Step 1: Prototype Selection**

1. **Feasibility** (2-3 years):

Developing a machine learning-driven framework for credit default prediction is feasible within the short-term future (2-3 years) for several reasons:

* **Availability of Data:** Financial institutions already possess extensive historical data, making it feasible to collect the necessary dataset for model development.
* **Advancements in Machine Learning**: Machine learning techniques, tools, and libraries have advanced significantly, making it easier to develop and deploy predictive models.
* **Compliance and Regulations**: Financial regulations and compliance standards are already in place, providing a framework for deploying predictive models in the industry.
* **Skilled Workforce**: There is a growing pool of data scientists and machine learning engineers with expertise in developing such models.

1. **Viability** (20-30 years):

The viability of a credit default prediction system over a 20-30 year timeframe is likely, but it depends on the evolution of the financial industry and technology:

* **Technological Advances**: If the system can adapt to future technological changes and integrate new data sources, it can remain viable.
* **Regulatory Changes**: Changes in financial regulations and privacy laws could impact the viability, but adaptable systems should remain relevant.
* **Market Dynamics**: The viability also depends on the persistence of the need for credit risk assessment, which is a fundamental aspect of the financial industry.

1. **Monetization**:

A machine learning-driven credit default prediction system can be directly monetized in the following ways:

* **Licensing**: Financial institutions can pay to license the technology or access the predictive model.
* **Subscription Model**: Offering a subscription-based service for ongoing access to credit risk assessments.
* **Consulting and Support**: Providing consulting and support services for implementing and fine-tuning the system within financial organizations.
* **Data Sales**: If the system generates valuable data, it can be monetized by selling anonymized, aggregated data to other businesses.

To enhance the long-term viability of the product/service, it's important to consider:

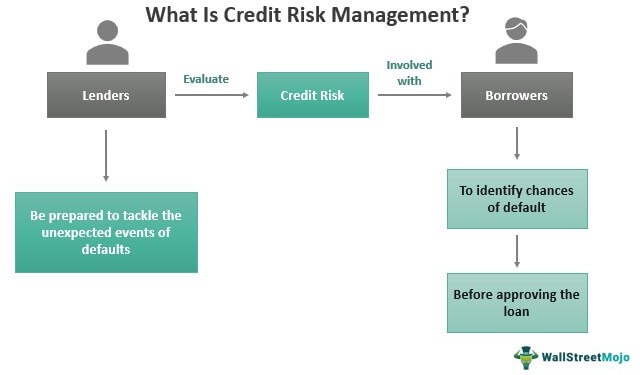
* **Scalability**: Ensure that the system can scale with growing data and user demands.
* **Adaptability**: The system should be designed to adapt to new data sources, changing regulations, and emerging technologies.
* **Ethical and Legal Considerations**: Keep a keen eye on ethical and legal aspects, especially with respect to data privacy and fairness.
* **Market Trends**: Continuously monitor financial industry trends and customer needs to remain relevant and competitive.

Ultimately, the feasibility, viability, and monetization potential of a machine learning-driven credit default prediction system depend on thorough planning, technology selection, and the ability to adapt to changes over time.

**Step 2: Prototype Development**

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.

**Step 3: Business Modelling**

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.

**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

**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 labelled 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

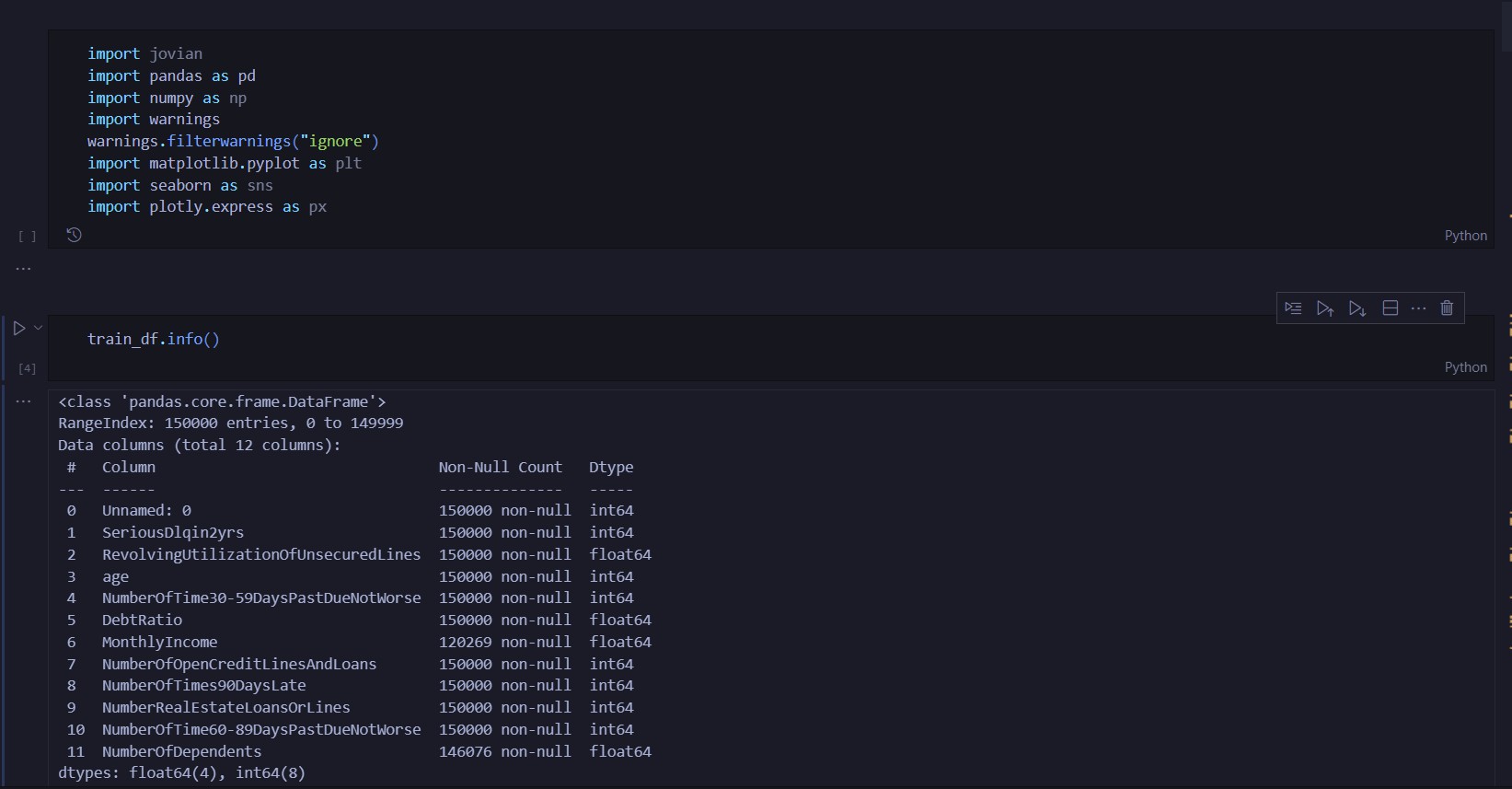
**What does it Cost?**

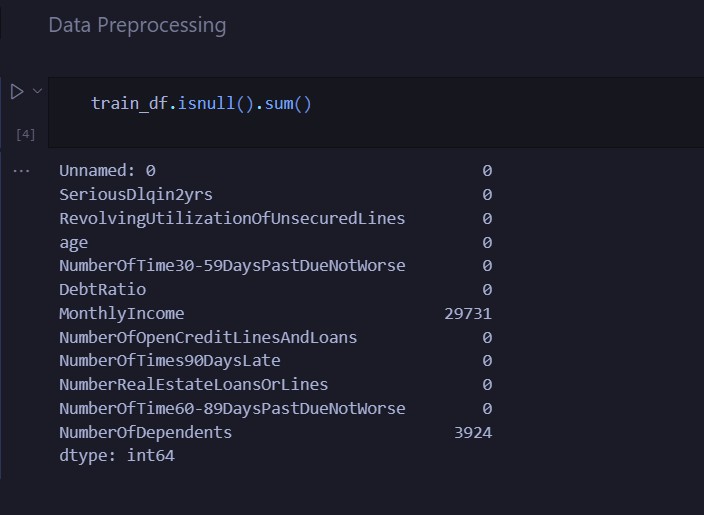
Developing and implementing a credit default prediction model involves several cost-driving factors that contribute to the overall expenses of the project:

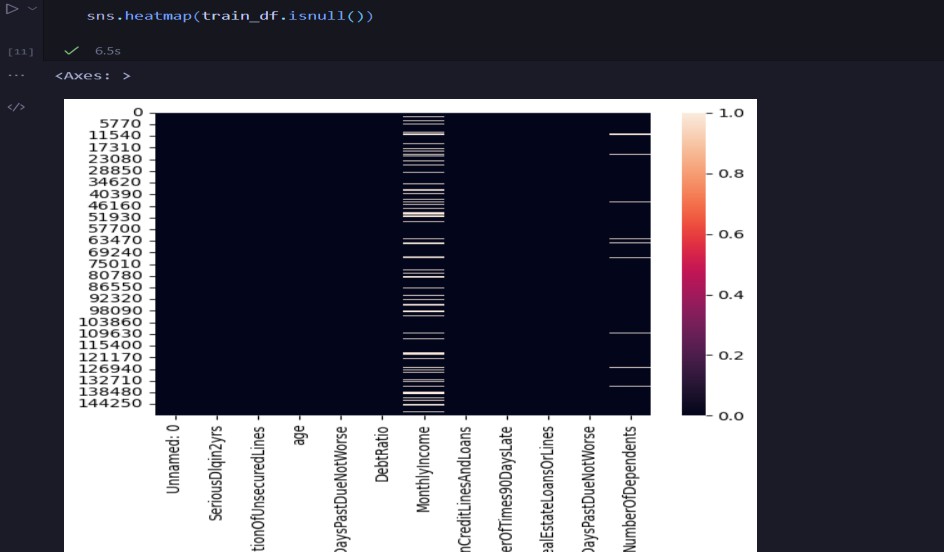
1. **Data Collection and Preprocessing:** Acquiring accurate and comprehensive credit- related data, including financial records and historical payment behaviour, might require partnerships with data providers or access to credit bureau databases, which can incur costs.
2. **Hard ware and software infrastructure:** Acquiring accurate and comprehensive credit-related data, including financial records and historical payment behaviour, 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.

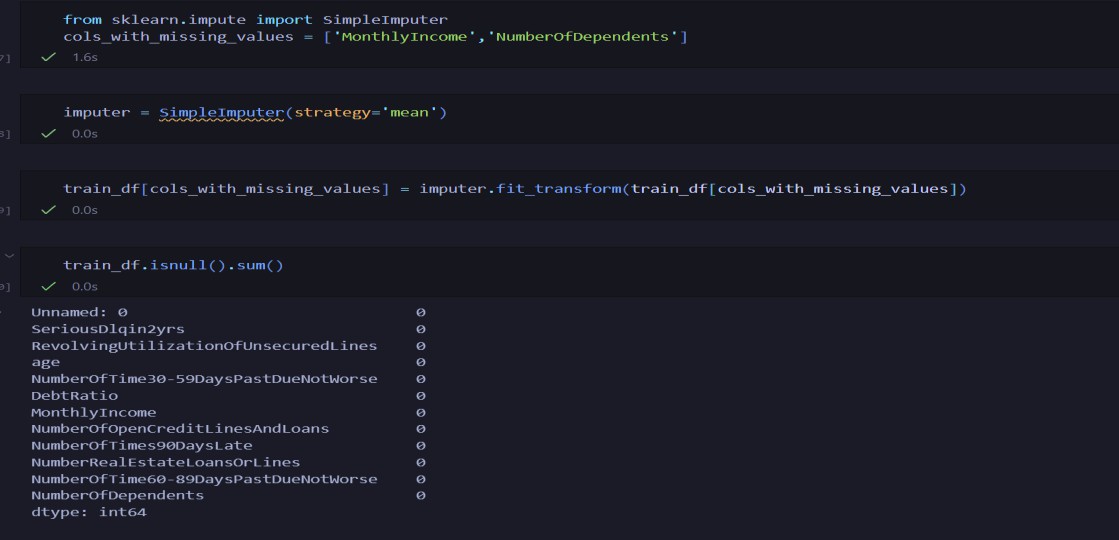
**Step 4: Financial Modelling (equation) with Machine Learning & Data Analysis**

**Product Details:**

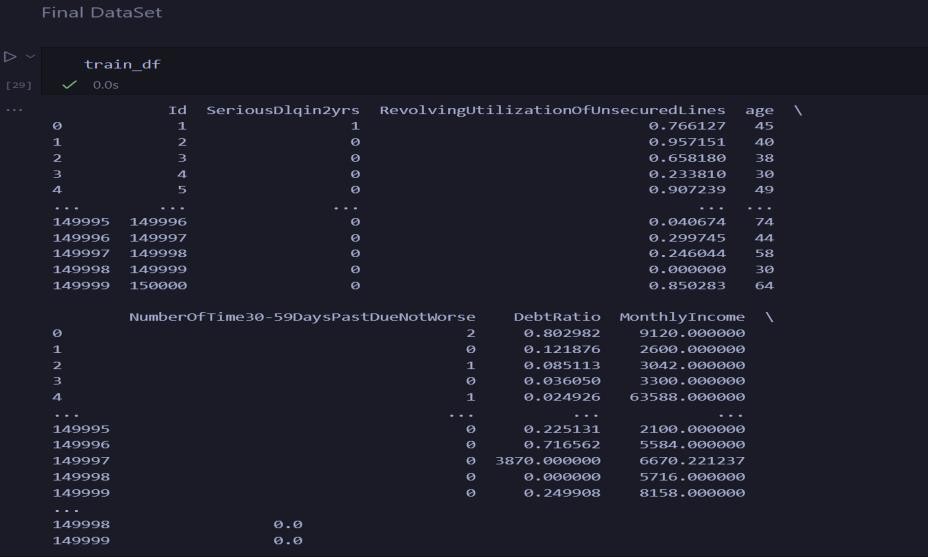
1. **Working:**
   1. **Data collection:** The initial step involves gathering a comprehensive and diverse dataset containing various factors associated with credit risk. This dataset encompasses historical credit data, financial attributes, borrower information, repayment history, and other relevant variables.
   2. **Data Preprocessing**: T 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 pre-processed 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:**
   1. **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.
   2. **Financial Institutions' Records:** Data from banks, lending companies, and credit card companies contain valuable information about borrowers' financial behaviors, loan applications, and repayment patterns.
   3. **Credit Bureaus:** Credit buxreaus compile credit reports containing credit scores, account histories, and credit inquiries. This data aids in assessing borrowers' creditworthiness
   4. **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. **Code Implementation:**
   * 1. Importing Libraries and Loading Dataset:
     2. Data Preprocessing:



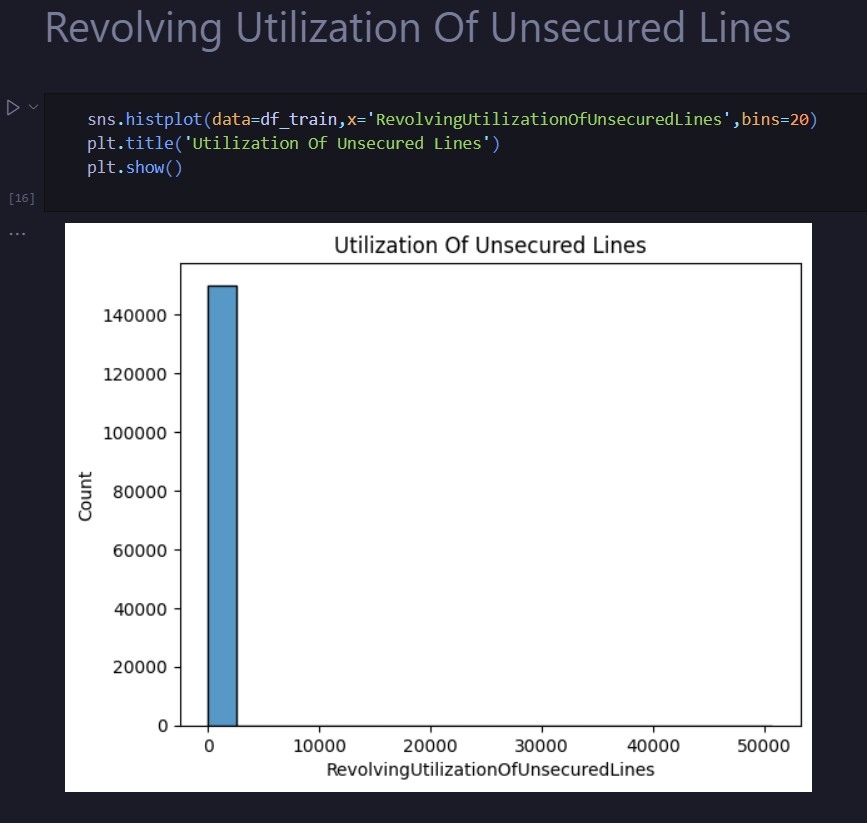


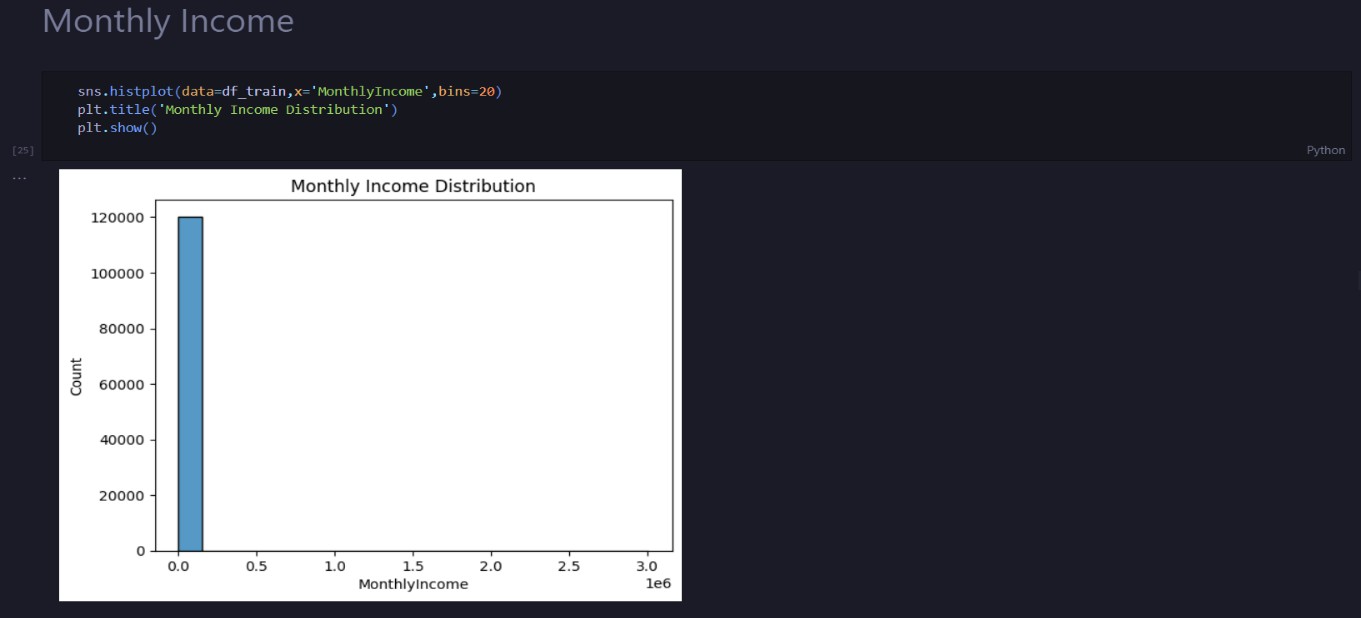
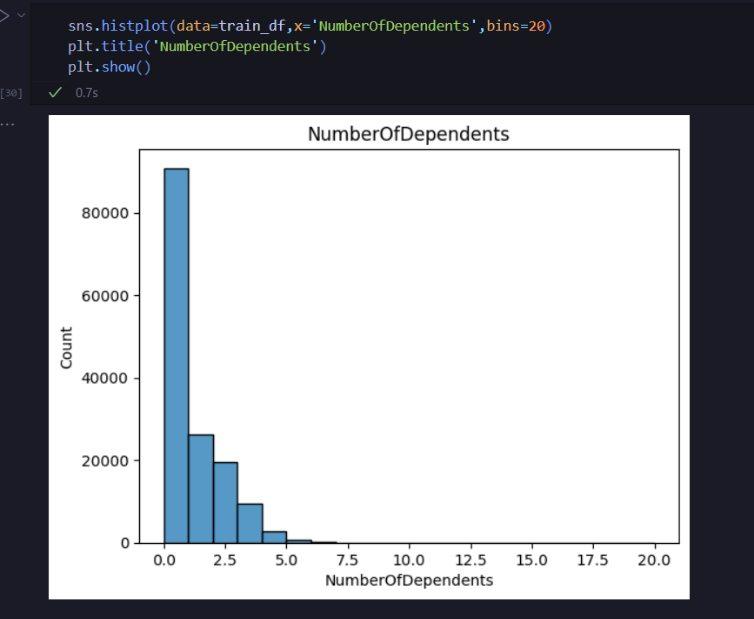


* + 1. Final Dataset:



* + 1. Exploratory Data Analysis:





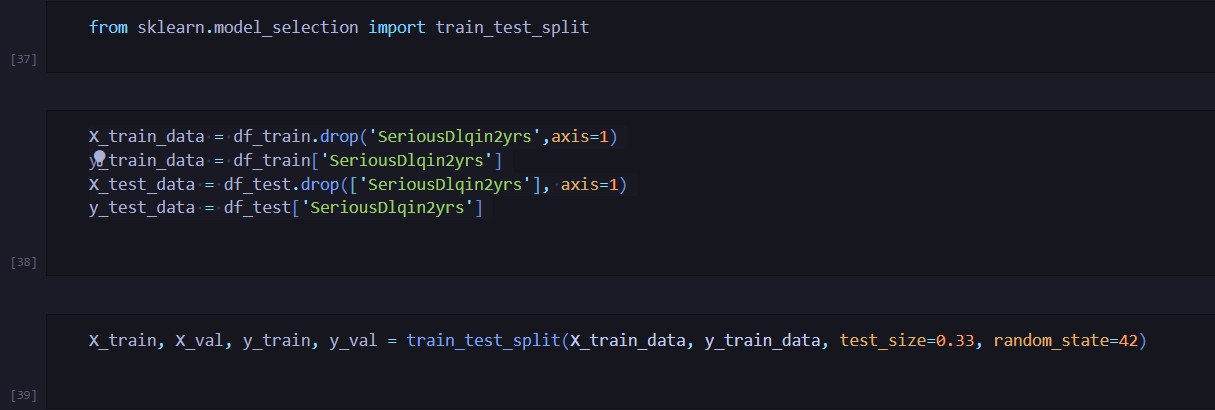
**Summary of EDA Findings:**

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.

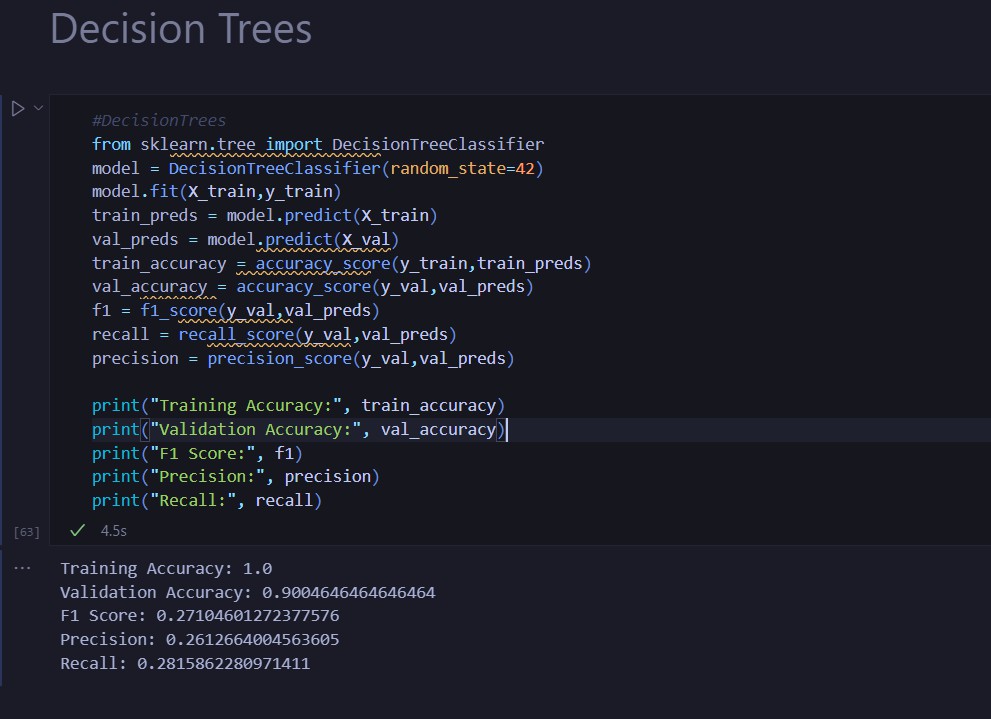
Complex Relationships and Missing Data: Key features like DebtRatio and MonthlyIncome show complex relationships and missing data points. High DebtRatio values sometimes coincide with missing MonthlyIncome, and the patterns are consistent in both the training and test datasets. Establishing a clear relationship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

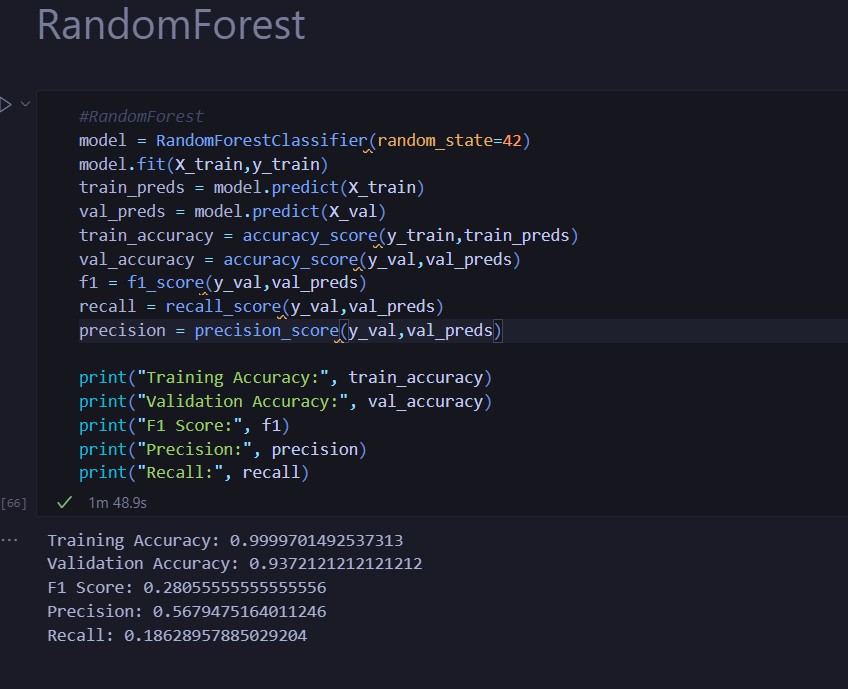
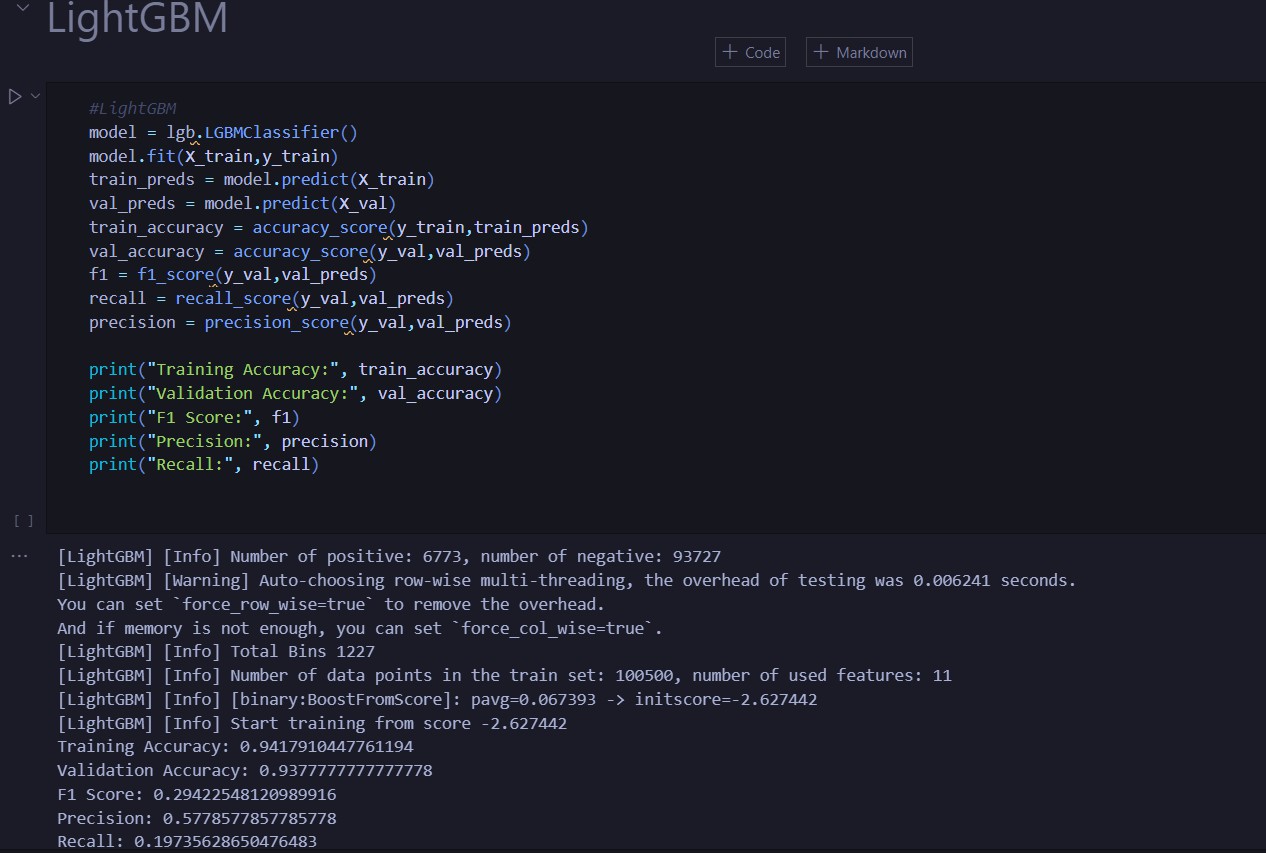
Credit Line and Loan Dynamics: The feature indicating the number of open credit lines and loans presents challenges in determining the distinction between sources of income and debts. While some instances involve implausible values or high numbers of dependents, filtering the data might not be appropriate, 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.

* + 1. Splitting the Dataset into training and validation sets:

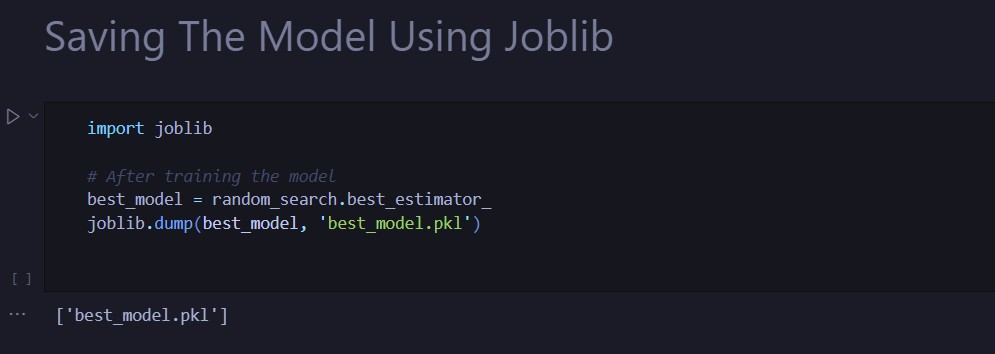


* + 1. Training of Machine Learning Algorithms:

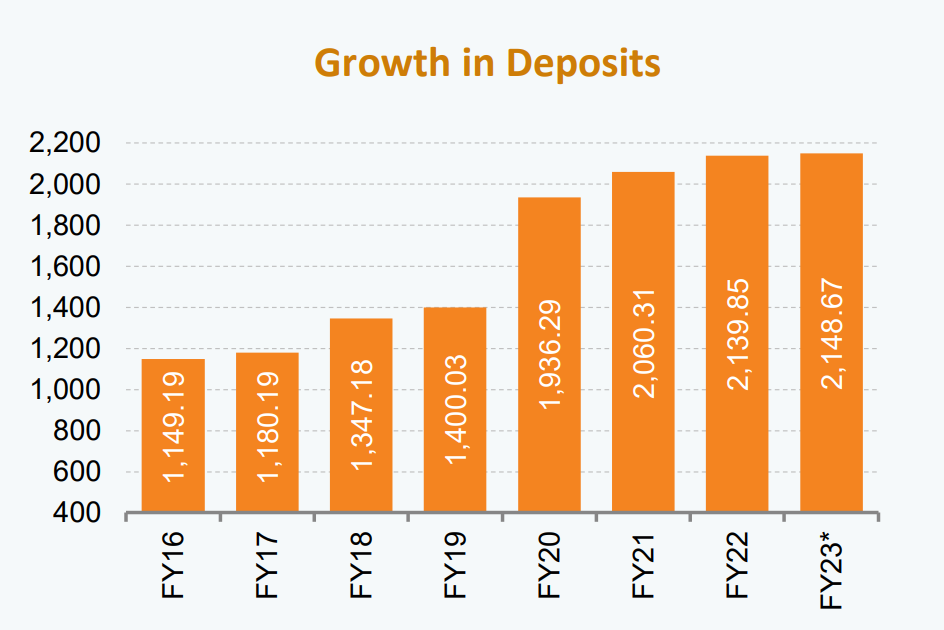


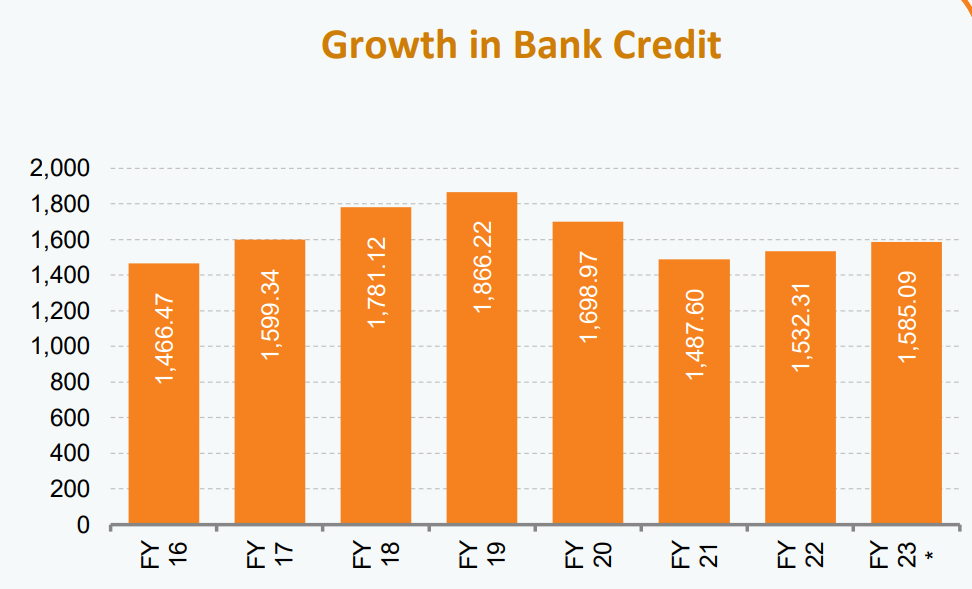


* + 1. Saving the model locally:



* + 1. Equation:





The above diagram shows the growth in Bank Credit for past 8 years. There was a negative impact of covid19, it has again started to grow. So if we assume the current trend to follow, we can charge for our service around Rs.80,000 on a monthly basis, we can increase the it further with more complexity and experience in future.

Let’s assume that the duration of developing the ML model takes about 2-6 months and the cost for producing the model is the salary of the members in the team including Data scientist and ML Engineers, Domain Experts, Data Analyst and Statisticians, Software Developers, Project Manages and Quality Assurance Specialist.

Let’s assume the total cost as C

So the Financial Equation will look like

Y = 80,000 X(t) – C

Here,,

X(t) = Function that represents the growth of the customer base

Y = Profit

# Future Update:

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

# 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.

<https://jovian.com/harshthakur22-7/credit-default-prediction-using-ml>