**CREDIT RISK ANALYSIS using PYSPARK**

A PROJECT REPORT

***Submitted by***

       BL.EN.U4AIE21015 B Sai Abhishek

       BL.EN.U4AIE21017 B Ruchith balaji

      BL.EN.U4AIE21038 Chillakuru Hari

**BACHELOR OF TECHNOLOGY**

IN

ARTIFICIAL INTELLIGENCE

***for the course***

***21AIE214-Big Data Analytics***

***Guided and evaluated by***

***Ms Niharika Panda Ma’am***

***Dept. of CSE,***



AMRITA SCHOOL OF COMPUTING, BANGALORE

 AMRITA VISHWA VIDYAPEETHAM

BANGALORE 560 035

June – 2023

**TABLE OF CONTENTS**

**Page no**

ACKNOWLEDGEMENTS 3

ABSTRACT 4

LIST OF FIGURES 5

CHAPTER 1- INTRODUCTION 6

* 1. Introduction to Credit Risk Analysis
  2. Motivation

CHAPTER 2 – LITERATURE SURVEY 8

CHAPTER 3 – SYSTEM DESIGN & IMPLEMENTATION 9

* 1. Logistic Regression
  2. Random Forest

CHAPTER 4 – RESULTS 15

CHAPTER 5 –CONCLUSION & FUTURE SCOPE 16

5.1 Conclusion

5.2 Future Scope

CHAPTER 6 – REERENCES 17

**Acknowledgement:**

*We offer our sincere pranams at the lotus feet of*

*Universal guru, MATA AMRITANANDA MAYI DEVI who showered her blessings throughout this project. We are very thankful to Br. Viswamrita Chaitanya Swamiji, Director, Amrita School of Engineering, Bangalore, for his valuable support. We are indebted to thank Dr. Sriram Devanathan, Principal, Amrita Vishwa Vidyapeetham, Bangalore for engraving a path for us to utilize the available resources to the fullest and there by widen our perspective of education and growth through it.*

*We deeply express our sincere thanks to Dr. Supriya, Chairperson, Department of Computer Science and Engineering for giving us the opportunity and encouraging us to use our abilities to the fullest in doing the project. It is our privilege to express our sincerest regards to our project guide, Ms. Niharika Panda, Lecturer, Department of Computer Science and Engineering, for the valuable inputs, guidance, encouragement, wholehearted cooperation and constructive criticism throughout the duration of our project.*

**Abstract:**

Credit risk analysis plays a crucial role in assessing the likelihood of borrowers defaulting on their loans, helping financial institutions make informed lending decisions and mitigate potential losses. With the increasing volume and complexity of credit data, traditional methods for credit risk analysis are often limited in their scalability and efficiency. In this project, we propose the implementation of credit risk analysis using PySpark, a powerful framework for distributed data processing and machine learning.Using PySpark, we leverage the parallel processing capabilities of Apache Spark to handle large-scale credit datasets efficiently. The project focuses on developing a credit risk model that incorporates various features, including demographic information, credit history, and financial indicators, to predict the probability of default. PySpark's distributed computing capabilities enable us to handle high-dimensional data and perform advanced data preprocessing tasks, such as feature scaling and missing value imputation.The credit risk analysis pipeline implemented in PySpark consists of several stages, including data ingestion, data preprocessing, feature engineering, model training, and model evaluation. PySpark's integration with machine learning libraries such as MLlib and MLflow enables us to utilize a range of classification algorithms, including logistic regression, random forests, and gradient boosting, to build robust credit risk models. Additionally, PySpark's scalability allows us to efficiently train models on large datasets and perform cross-validation for model selection and hyperparameter tuning.

**LIST OF FIGURES**

*Fig-1 Logistic Regression Model Example*

*Fig-2 Random Forest Model Example*

*Fig-3 Correlation Matrix Relation b/w every Column*

*Fig-4 Logistic Model Design of LR model*

*Fig-5 Random Forest Model Design of RF model*

*Fig-6 LR-Result Prediction*

*Fig-7 RF-Result Prediction*

**CHAPTER - 1**

**INTRODUCTION**

**Introduction to Credit Risk Analysis:**

Credit risk analysis is a vital task for financial institutions to assess the probability of borrowers defaulting on their loans. Accurate credit risk assessment enables lenders to make informed decisions, manage their portfolios effectively, and mitigate potential losses. However, with the increasing volume and complexity of credit data, traditional methods for credit risk analysis face challenges in terms of scalability and efficiency.

To address these challenges, this project proposes the implementation of credit risk analysis using PySpark, a powerful framework for distributed data processing and machine learning. PySpark leverages the parallel processing capabilities of Apache Spark to handle large-scale credit datasets efficiently. By employing PySpark, financial institutions can benefit from the scalability and speed of distributed computing, enabling them to analyze vast amounts of credit data in a timely manner.

The objective of this project is to develop a credit risk model that integrates various features, such as demographic information, credit history, and financial indicators, to predict the probability of default. PySpark provides a comprehensive set of tools for data preprocessing, feature engineering, model training, and evaluation, making it an ideal choice for building and deploying robust credit risk models.

The benefits of implementing credit risk analysis using PySpark are significant. Financial institutions can leverage the scalability and distributed computing capabilities of PySpark to process large volumes of credit data, gain valuable insights, and make informed credit decisions. Additionally, PySpark's flexibility allows for easy integration with existing data processing pipelines, ensuring a seamless transition for organizations adopting this approach.

In this project, we aim to demonstrate the effectiveness and efficiency of credit risk analysis using PySpark. By leveraging the power of distributed computing, we can enhance the accuracy and speed of credit risk assessments, leading to improved risk management and decision-making in the lending industry.

**Motivation:**

The motivation for credit risk analysis stems from the need for financial institutions to effectively evaluate and manage the risks associated with lending money. Several key factors drive the importance of credit risk analysis:

***Mitigating potential losses*:** Lending inherently involves the risk of borrowers defaulting on their loans, which can result in financial losses for lenders. By conducting thorough credit risk analysis, financial institutions can assess the likelihood of default and take proactive measures to mitigate potential losses.

***Enhancing decision-making:*** Credit risk analysis provides valuable insights into borrowers' creditworthiness, enabling financial institutions to make informed lending decisions. By leveraging data-driven models and analytics, lenders can improve the accuracy of credit risk assessments and enhance their decision-making processes.

***Strengthening customer relationships:*** Sound credit risk analysis practices help financial institutions build trust with their customers. By accurately assessing credit risk and tailoring loan terms accordingly, lenders can offer competitive interest rates and favorable terms, strengthening customer relationships and loyalty.

**CHAPTER – 2**

**LITERATURE SURVEY**

**Paper - [1]: View of Evaluating Classical and Artificial Intelligence Methods for Credit Risk Analysis**

Author research allowed for the comparison of statistical and Al predictors, adding significantly to the academic literature by designing a credit scoring experiment that compares various distinct types of models using a novel dataset with financial and other relevant data for a selection of Portuguese companies. Credit scoring methods were successfully implemented based on this information and used to distinguish between good and bad applicants in the timespan of a year.

Now we had found that their Model/ Research has some gaps that when we tested the model, It is Less efficient and less predictive. So we had decided to fill that gap.

We are Going to use Pyspark with Logistic Regression and Random Forest model to create correct predicting Model.

**Paper – [2]:**

Author described machine learning approaches to assess credit candidate applicants’ profiles and continued credit scoring based on the non-traditional dataset of Lending Club Company. Regarding the classification accuracy, the results showed that the KNN is more accurate, informative, and conservative for personal credit evaluation. Furthermore, the model predictions could be used to score new and old clients in an accurate scorecard complementary to traditional credit evaluation methods.

In above study they have used pyspark with KNN. When we tested the results showed that the model is predicting the output slow. We have decided to improve the Model by using Logistic Regression and Random Forest.

**CHAPTER – 3**

**SYSTEM DESIGN & IMPLEMENTATION**

In this project we are going to Implement Credit Risk Analysis using Pyspark.

**Brief:**

**Pyspark:**

PySpark is a powerful framework for distributed data processing and machine learning, making it well-suited for credit risk analysis.

Pyspark can be helpful for following:

1. Scalability and Distributed Computing.
2. Data Preprocessing.
3. Feature Engineering
4. Machine Learning Algorithms
5. Model Evaluation and Selection
6. Real-Time Analysis
7. Scalable Model Deployment

PySpark provides a comprehensive set of machine learning algorithms through its MLlib and ML libraries. Here is a list of commonly used machine learning models available in PySpark

**Linear Regression:** A linear model used for regression tasks to predict continuous numeric values.

**Logistic Regression:** A classification model used to predict binary outcomes or perform binary classification tasks.

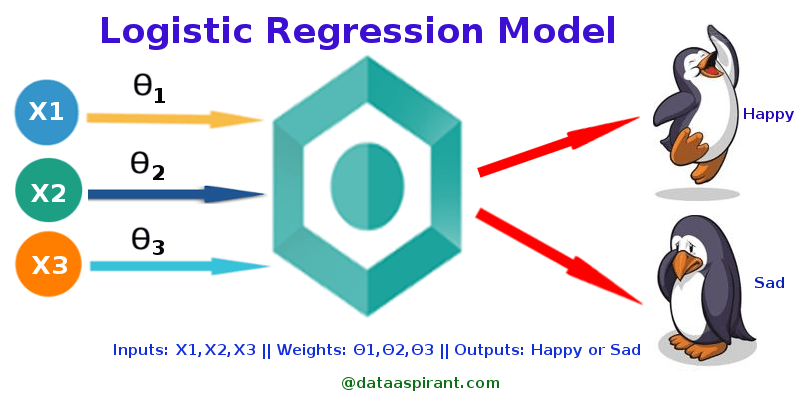
**Decision Trees:** A non-parametric supervised learning model that creates a tree-like model of decisions based on feature values.

**Random Forest:** An ensemble learning method that combines multiple decision trees to improve predictive accuracy and reduce overfitting.

**Gradient-Boosted Trees (GBT):** A boosting algorithm that builds an ensemble of decision trees sequentially, with each tree attempting to correct the mistakes of the previous ones.

Among the above five we are gong to use **Logistic Regression** and **Random Forest** in the presence of **Pyspark** to implement Best Machine Learning Model to predict Credit Risk.

**Logistic Regression:**



*Fig -1 Logistic Regression*

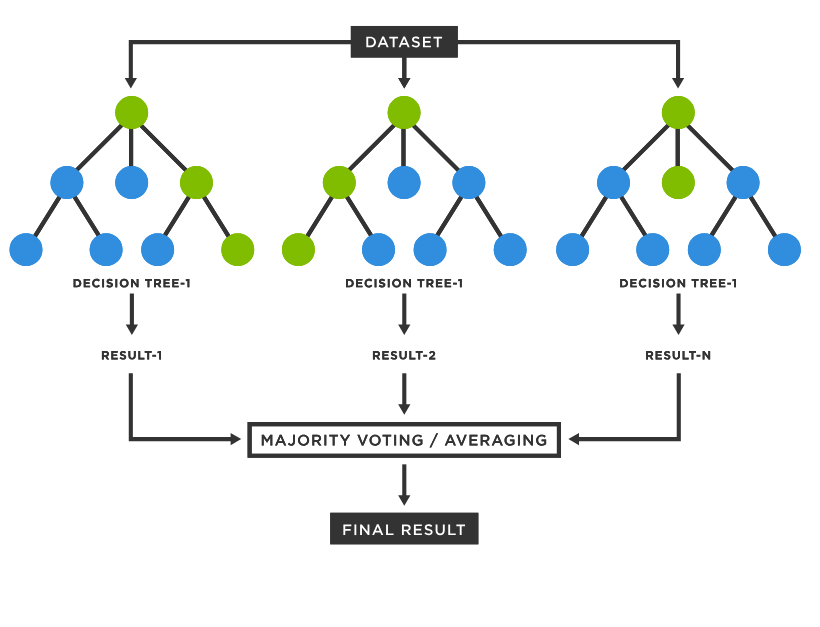
Logistic regression is a widely used statistical model for binary classification tasks. It is a supervised learning algorithm that predicts the probability of an event or outcome belonging to a specific class. Although the name "regression" is used, logistic regression is a classification algorithm rather than a regression algorithm.

**Binary Classification:** Logistic regression is primarily used for binary classification problems where the target variable takes one of two possible classes, such as "yes" or "no," "fraud" or "not fraud," or "default" or "non-default."

**Model Evaluation:** Logistic regression models can be evaluated using various metrics such as accuracy, precision, recall, F1 score, and area under the Receiver Operating Characteristic (ROC) curve. These metrics assess the performance of the model in terms of its predictive accuracy and ability to correctly classify instances.

Logistic regression is widely used in various domains, including finance, healthcare, marketing, and fraud detection. Its interpretability, simplicity, and effectiveness make it a popular choice for binary classification tasks. In PySpark, logistic regression can be implemented using the MLlib or ML libraries, allowing for distributed processing and scalability.

**Random Forest:**

****

*Fig-2 Random Forest*

Random Forest is an ensemble learning method that combines multiple decision trees to improve predictive accuracy and reduce overfitting. It is a popular machine learning algorithm available in PySpark's MLlib and ML libraries.

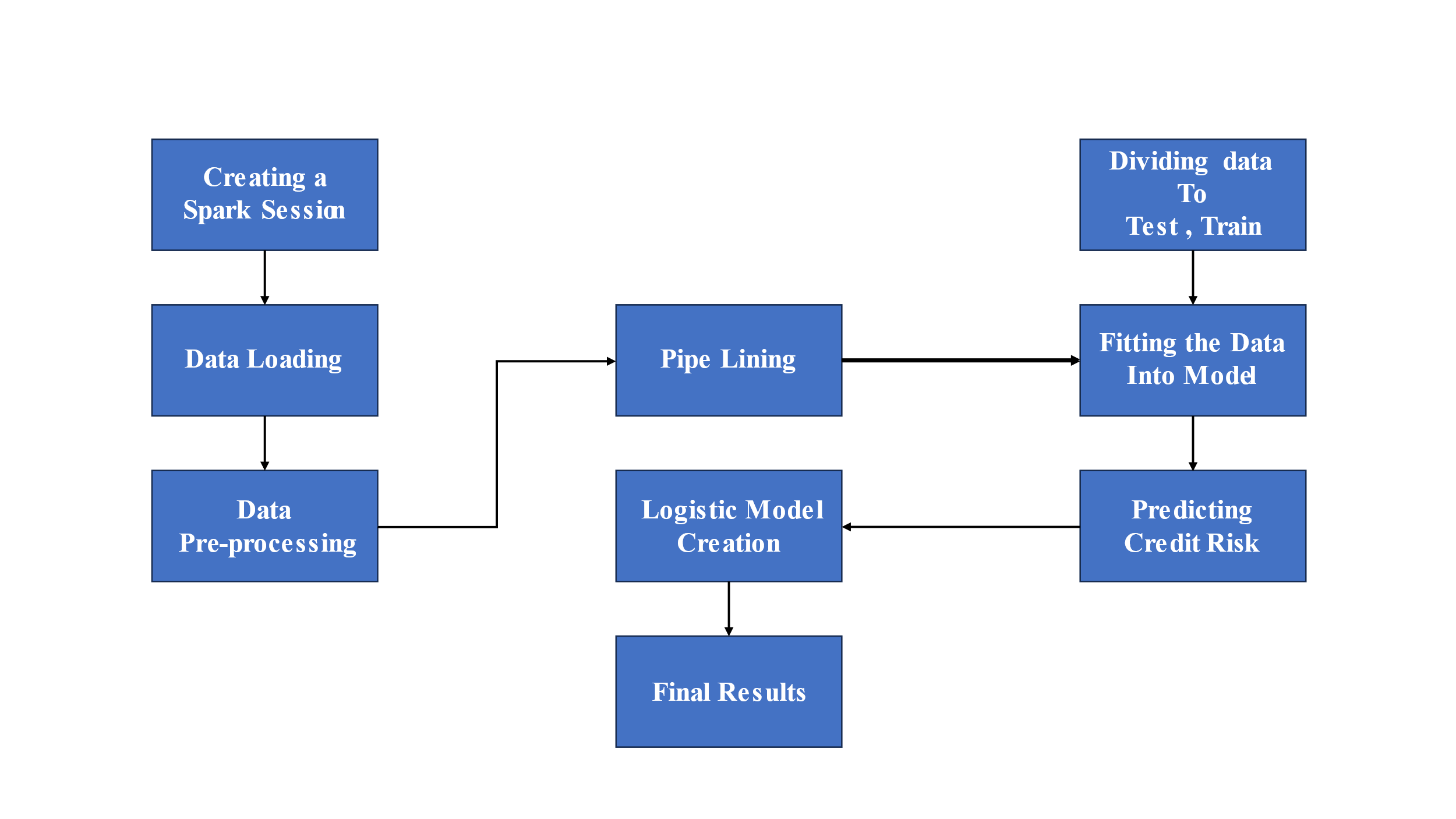
Ensemble Learning: Random Forest is based on the concept of ensemble learning, which combines the predictions of multiple models to make more accurate and robust predictions. In the case of Random Forest, the ensemble consists of multiple decision trees.

**Voting Mechanism:** During prediction, Random Forest aggregates the predictions of all the individual trees and outputs the majority prediction (for classification) or the average prediction (for regression). This voting mechanism helps make the final prediction more accurate.

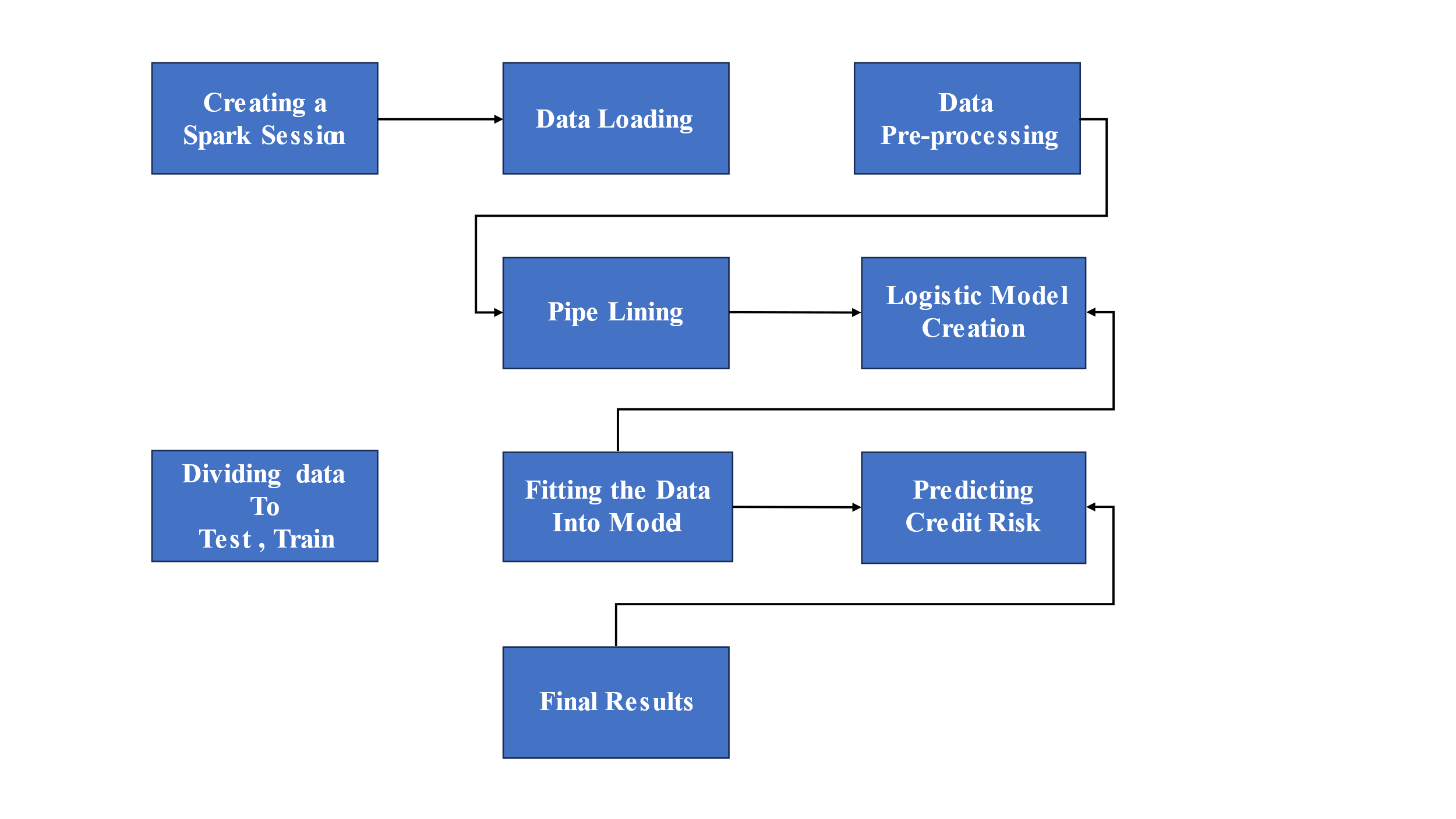
**Feature Importance:** Random Forest provides a measure of feature importance, which indicates the relative contribution of each feature in the model's decision-making process. Feature importance can help identify the most influential features and provide insights into the underlying patterns and relationships in the data.

**Handling Missing Data and Outliers:** Random Forest can handle missing data and outliers effectively. It leverages the ensemble of trees to make predictions based on the available information in the other features. This makes Random Forest more robust to missing data and less prone to the influence of outliers compared to individual decision trees.

**Design:**

**For Logistic Regression:**

*Fig – 3 LR Model Design*

**For Random Forest:**

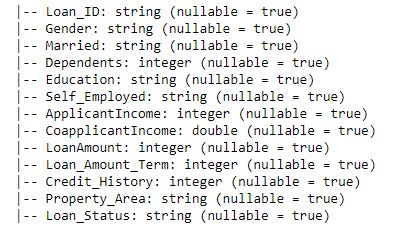
*Fig -4 RF Model Design*

**Implementation:**

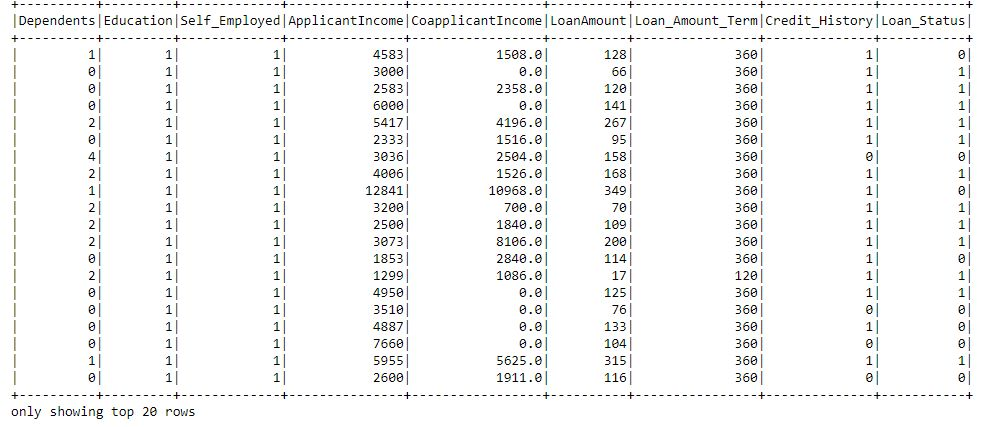
1. **Data Set Loading:**

The data set loaded into the Model using Spark.read

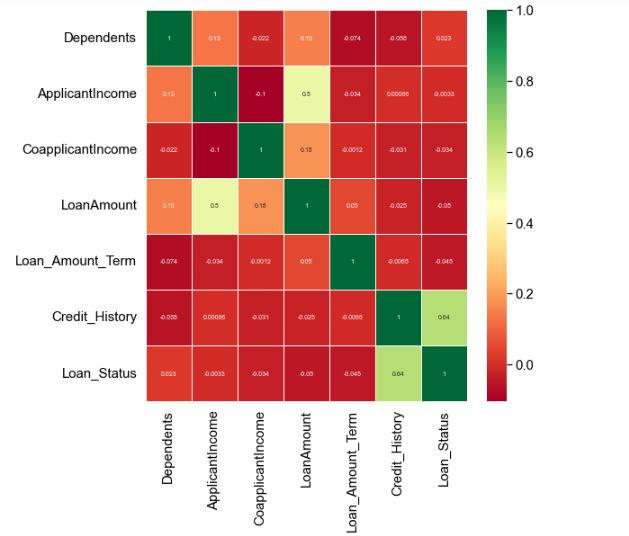
The Data set have the following Columns with its Data types:

****

1. **Preprocessing:**

****

The data have Missing values Wrong Values and String Values which Causes Error to the Data.

1. ******Correlation Matrix:**

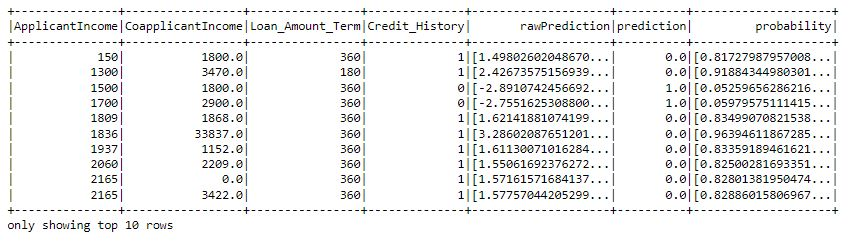
*Fig-5 Correlation Matrix*

**CHAPTER – 4**

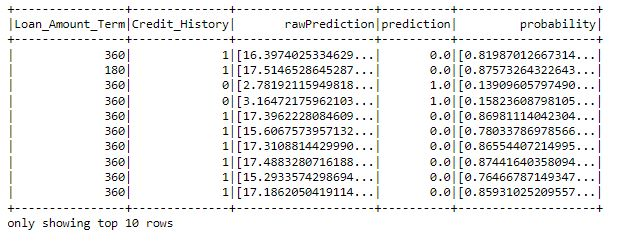
**RESULTS AND ANALYSIS**

**Result:**

1. **Logistic Regression:**

*****Fig-6 LR Result*

1. **Random Forest:**

****

*Fig -7 RF Result*

The Result table has Prediction and Probability columns Which show whether to approve Loan or not to the Customer.

**CHAPTER – 5**

**CONCLUSION AND FUTURE SCOPE**

**Conclusion:**

In conclusion, the implementation of credit risk analysis using PySpark provides financial institutions with a powerful and scalable framework for analyzing credit risk. By leveraging PySpark's distributed computing capabilities, data preprocessing functions, feature engineering tools, and a wide range of machine learning models, financial institutions can make accurate and timely credit risk assessments, leading to improved risk management and decision-making processes.

**Future Scope:**

Incorporating advanced machine learning techniques can enhance credit risk analysis. PySpark can leverage state-of-the-art algorithms like deep learning, gradient boosting, and natural language processing for more accurate risk assessment. These techniques can capture complex patterns and relationships in the data, leading to improved predictive models.Real-time credit risk analysis is gaining importance as financial institutions aim to make instant decisions. PySpark's integration with Apache Spark Streaming enables continuous monitoring and analysis of streaming credit data. This allows for timely detection of credit risk events and proactive risk management.

**REFERENCES**

1. C. Onay and E. Ozt¨urk ¨ A review of credit scoring research in the age of Big Data J.

Financ. Regul. Compliance, vol. 26, no. 3, pp. 382–405, Jul. 2018.

2. A. AshoftehMining Big Data in statistical systems of the monetary financial institutions

(MFIs) in International Conference on Advanced Research Methods and Analytics (CARMA),

2018, p. doi: 10.4995/carma2018.2018.8570.

3. M. Oskarsd´ottir, C. Bravo, C. Sarraute, J. Vanthienen, and B. Baesens ´ The value of

big data for credit scoring: Enhancing financial inclusion using mobile phone data

and social network analytics Appl. Soft Comput. J., vol. 74, pp. 26–39, Jan. 2019.

4. J. S. Pedro, D. Proserpio, and N. OliverMobiscore: Towards universal credit scoring

from mobile phone data in Lecture Notes in Computer Science (including subseries

Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2015,

vol. 9146, pp. 195–207.