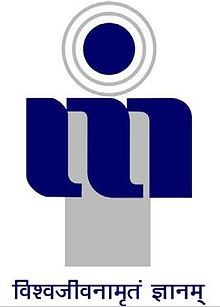
****

**ABV-Indian Institute of Information Technology & Management, Gwalior**

*Project Report*

*of*

*AI Workshop*

**Credit Card Fraud Detection Using Machine Learning**

**Members:**

*Amgoth Sai Teja - 2021IMG007*

*Anushka Singh - 2021IMG015*

*Lives kumar Singh - 2021IMG032*

*Rupali Das - 2021IMG051*

**Under the Supervision Of:**Dr Veena Anand

**Report Certificate**

I hereby certify that the work, which is being presented in the report entitled Student’s Performance Analysis, Bachelor of Technology in IPG-MBA and submitted to the institution is an authentic record of my/our work carried out during the period October 2023 under the supervision of course conductors. I also provided the citation for the text, figure, and table they were pulled from.

Course Conductors:

Dr Veena Anand

Date: 01/11/2023

The signatories have examined the ﬁnal copy of this report, and we ﬁnd that both the content and the form meet acceptable presentation standards of scholarly work in the discipline mentioned above.

**Acknowledgement**

We want to express our sincere gratitude and appreciation to all those who contributed to the successful completion of our Credit Card Fraud Detection project, which was implemented using logistic regression. This project would not have been possible without the invaluable support and assistance of numerous individuals and organizations, and we would like to acknowledge their contributions.

First and foremost, we extend our heartfelt thanks to our project supervisor, Dr Veena Anand, for their guidance, mentorship, and continuous support throughout the project. Their expertise in data analysis and machine learning was instrumental in shaping the direction of our research and ensuring its success.

We would like to acknowledge our institution, ABV-IIITM Gwalior, for providing us with the necessary resources, access to data, and a conducive research environment. The academic community and the library staff were crucial in facilitating our data collection and analysis.

Special thanks go to our fellow team members for their dedication and collaborative efforts. Each team member contributed their unique skills and expertise, which greatly enhanced the quality of our project. Their commitment and hard work were instrumental in overcoming the challenges we faced.

We would also like to express our gratitude to the website Kaggle that provided us with the dataset used in this project. Without access to this critical data, our research would not have been possible.

**Abstract**

In the financial sector, credit card fraud is a widespread and expensive issue, resulting in significant losses for financial institutions and consumers. The ever-evolving fraudsters' strategies frequently make it difficult for conventional rule-based fraud detection systems to stay up. To solve this problem, we suggest a sophisticated Credit Card Fraud Detection System (CCFDS) based on artificial intelligence (AI) methods.

The increased use of online transactions has resulted in an alarming surge in credit card fraud, resulting in considerable financial losses for both people and financial institutions. We offer an AI-based Credit Card Fraud Detection System that employs cutting-edge machine learning and deep learning algorithms to solve this essential issue. This project aims to provide a robust and efficient solution to detect fraudulent credit card transactions in real time.

The business's biggest obstacle is that neither the card nor the cardholder must be present while making the purchase. As a result, the merchant needs help determining whether or not the consumer making a purchase is a legitimate cardholder. Analyzing data sets and user-current datasets through classification. The correctness of the outcome data should then be optimised. The procedures' accuracy, sensitivity, specificity, and precision are used to evaluate their performance. The evaluation of some of the attributes provided then finds fraud detection.

**Dedication**

The dedication necessary to construct an AI/ML project on credit card fraud detection originates from the crucial relevance of addressing financial security concerns and protecting consumers, organizations, and financial institutions from fraudulent actions. It is motivated by the critical need to improve financial security, safeguard persons, and assure the dependability of financial transactions. This commitment incorporates technological, ethical, and societal issues, advocating ongoing progress and innovation in the sector in order to remain ahead of fraudulent actions and maintain financial integrity.

This fact has motivated me to study about various kinds of fraud and illegitimate activities that can occur in case of credit card financial transactions and motivated me to build a project which can detect the same.

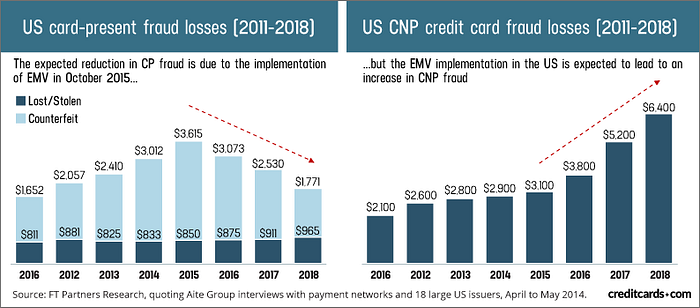
This project uses the following technology and tools:

* Flask for implementing and developing in Python.
* Numpy and Pandas for data cleaning.
* Pickle to serialize a python object into binary format and vice versa.
* Google Colab to run code entirely on cloud.
* Learn to build and implement models.

**Introduction**

Artificial intelligence (AI) and machine learning (ML) technologies have revolutionized various industries in the current digital era. These technologies offer solutions to intricate problems and have changed the way data analysis and decision-making processes are approached. This report presents an in-depth exploration of an AI/ML project dedicated to credit card fraud detection.

The widespread use of credit card transactions has provided consumers and businesses with unmatched convenience in an increasingly linked digital economy. But along with its convenience comes a rising threat in the shape of fraudulent activity. This study summarizes a painstakingly planned AI/ML project that tackles the crucial problem of credit card fraud detection. This project's goal is to create a reliable and flexible system that uses machine learning (ML) and artificial intelligence (AI) to identify and stop fraudulent credit card transactions.



Fighting credit card theft is important since it affects both the economy and society as a whole. Consumer confidence in the security of electronic transactions is undermined by fraudulent operations, which also cause significant financial losses. Through strengthening security protocols, lowering financial risks, and maintaining the dependability of credit card transactions, this project has the potential to make a substantial contribution.

**Important Terms**

1. **Artificial Intelligence**

Artificial Intelligence (AI) is the pursuit of intelligent systems capable of imitating human-like cognitive functions. It entails creating software, algorithms, and other technological advancements that enable robots to carry out operations that ordinarily call for human intellect. Enabling these systems to think, learn, see, and act like humans while maintaining computing efficiency and accuracy is the main objective.

Machine learning is one of the more well-known subfields within artificial intelligence (AI). Without explicit programming, machines may learn from data, spot patterns, and make judgments or predictions thanks to machine learning algorithms. Machines can now read, interpret, and produce human language thanks to natural language processing (NLP), while computer vision makes it easier to understand images and videos.

Artificial Intelligence has several applications that affect many different industries. AI helps with drug discovery, personalized treatment, and diagnostics in the healthcare industry. It improves trading tactics, risk evaluation, and fraud detection in the financial industry. AI is beneficial to transportation since it allows for traffic planning and driverless cars. AI is used in entertainment to create personalized content recommendations. The world we live in is constantly changing because to AI. Fairness, accountability, and transparency are ethical issues in AI that become increasingly important as AI systems have an impact on decision-making. The effects of AI on the workforce provoke debates about the necessity for upskilling and the effects of automation on employment.

1. **Machine Learning**

A branch of artificial intelligence called machine learning (ML) focuses on developing models and algorithms that let computers learn from data and make judgments or predictions without the need for explicit programming. ML algorithms enable computers to recognize patterns, gain insights, and enhance performance via experience by utilizing statistical techniques. Large-scale data sets are used to train models, which enables systems to identify patterns and produce precise predictions or classifications. This is the fundamental idea of machine learning.

While unsupervised learning finds patterns in unlabeled data, supervised learning uses labeled data to train models to make predictions. Systems learn through interactions with their surroundings and feedback in the form of rewards or penalties, which is known as reinforcement learning. ML has wide-ranging applications in many different fields. It helps with diagnosis and individualized treatment plans in the medical field. It drives fraud detection and risk assessment in the financial sector.

Recommendation systems are useful in e-commerce, and predictive maintenance minimizes downtime in production. Deep learning, a branch of machine learning that uses neural networks to improve speech recognition, picture classification, and natural language processing, is an ongoing development in the field. Data quality, the interpretability of intricate models, and the moral issues of bias and justice in algorithms are ongoing concerns. The field of machine learning is changing, bringing with it both opportunities and challenges that will change the way we solve issues and use technology.

1. **Data Science**

Data science is an interdisciplinary field that uses systems, algorithms, and scientific methods to combine to extract useful knowledge and insights from data. To arrive at relevant conclusions it entails gathering, preprocessing, analyzing, interpreting, and visualizing data—both structured and unstructured. Fundamentally, data science involves multiple phases. Data must first be gathered and acquired from a variety of sources. Then, raw data must be cleaned, arranged, and transformed into an organized format that can be analyzed.

In order to find patterns or trends, exploratory data analysis, or EDA, involves summarizing data and presenting it visually. To analyze data, predict trends, or make predictions, data scientists use a variety of modeling techniques, including statistical models and machine learning algorithms. Large volumes of data can be handled and processed effectively with the help of big data technologies like Hadoop and Spark.

Data science requires careful consideration of ethical issues pertaining to security, fairness, and privacy of data. The field has applications in marketing, finance, healthcare, and other areas. It promotes innovation, process optimization, and well-informed decision-making. Data science is always evolving due to ongoing advancements in techniques and tools, which present both new challenges and opportunities in a world where data is becoming more and more important.

1. **Regression**

The statistical technique of regression is employed to examine and simulate the correlation between a dependent variable and one or more independent variables. Based on the values of the independent variables, it seeks to forecast or explain the behavior of the dependent variable. The most popular type is called linear regression, in which a linear equation represents the relationship between the variables. One independent variable is used to predict the dependent variable in simple linear regression. Several independent variables can affect the dependent variable in multiple linear regression.

Non-linear regression techniques, such as logistic regression or polynomial regression, deal with non-linear relationships. Regression analysis entails fitting a mathematical model to empirical data. It makes predictions or explains the relationship between variables by estimating the model's parameters to best fit the data. Regression analysis is used to forecast trends, explore relationships, and make estimations based on available data in a variety of fields, including engineering, economics, social sciences, and healthcare.

By penalizing extreme parameter values, regularized regression techniques like ridge and lasso regression impose constraints to either improve model accuracy or prevent overfitting. Bayesian and time-series regression are two other specialized techniques that are tailored to particular data types or predictive tasks.

1. **Data and Preprocessing**

**5.1 Dataset**

A dataset is a fundamental element in the field of data and preprocessing in artificial intelligence and machine learning. It is made up of an organized or unorganized collection of data that is used to train and evaluate machine learning models. This dataset consists of discrete data points, each of which has a collection of characteristics or features that define the observations.

A crucial step in AI/ML is data preprocessing, which entails organizing and improving this dataset to improve its quality and usefulness. This includes activities like encoding categorical variables, handling missing values, dealing with outliers, standardizing or normalizing numerical features, and dividing the dataset into training and testing sets.

The dataset is improved and optimized for machine learning algorithms through these preprocessing stages, guaranteeing the models' precision, effectiveness, and capacity to identify significant patterns or insights throughout the training phase.

**5.2 Feature**

A feature is an individual attribute or characteristic within a dataset that represents particular aspects or measurements associated with each data point in data and preprocessing within AI/ML. Features, which provide crucial information that machine learning models use to recognize patterns and relationships, can be textual, numeric, or categorical.

A crucial part of data preprocessing is feature engineering, which is choosing, modifying, or generating features to improve model performance. While feature creation or transformation refines data to help algorithms make more accurate predictions or classifications, effective feature selection lowers noise and increases model accuracy. The dataset is made more machine learning algorithm-friendly by optimizing its features, which promotes improved analysis and decision-making.

**5.3 Preprocessing**

The preparatory actions carried out on raw data prior to using it in machine learning models are referred to as preprocessing in data and preprocessing within AI/ML. This crucial phase includes data organization, transformation, and cleaning. These procedures involve categorical variable encoding, addressing outliers, scaling or normalizing features, and handling missing values.

Preprocessing's main goal is to improve the data so that machine learning algorithms can use it. A well-preprocessed dataset guarantees that the data is complete, consistent, and formatted optimally for efficient analysis and model training, which improves the models' accuracy and performance.

1. **Machine Learning Models**
   1. **Supervised Learning**

A key idea in machine learning is supervised learning, in which an algorithm gains knowledge from labeled training data in order to classify or predict. The training dataset, which provides both the matching accurate output and the input data, is where the "supervision" originates. The algorithm's main objective is to figure out the relationship or mapping between the input and output data.

In supervised learning, algorithms acquire the ability to extrapolate patterns from labeled data in order to forecast on future data using:

* ***Regression:*** The method forecasts a continuous output in regression tasks. Predicting a house's price, for instance, depending on its attributes, such as size, location, number of bedrooms, etc.
* ***Classification:*** The algorithm predicts a category output in classification problems. This can involve multi-class classification, which sorts data into more than two classifications, such as fruit varieties, or binary classification, which divides data into two classes, such as spam or non-spam.

In order to facilitate automation and decision-making based on learned patterns from data, supervised learning is a potent and extensively utilised machine learning technique. It is employed in various fields, including image identification, natural language processing, recommendation systems, healthcare, finance, and more.

* 1. **Unsupervised Learning**

Unsupervised learning is a subfield of machine learning that aims to discover structures or patterns in unlabeled data. The algorithm cannot predict or learn from specific labels or predetermined outcomes, in contrast to supervised learning. Rather, the algorithm itself investigates the data to find hidden correlations, patterns, or structures. The goal of unsupervised learning is to identify innate structures or correlations in a dataset by using input data without any labeled answers. When the goal is to investigate and comprehend the underlying features or structure of the data, this kind of learning is especially helpful.

There are two main categories of unsupervised learning: dimensionality reduction and clustering:

* In unsupervised learning, ***clustering*** is a popular job in which algorithms combine comparable data points according to their intrinsic properties.
* ***Dimensionality reduction techniques*** are also an important component of unsupervised learning. By lowering the amount of features or variables and keeping as much pertinent information as feasible, they seek to simplify the data.

Unsupervised learning methods play a crucial role various fields, including anomaly detection in cybersecurity, customer segmentation in marketing, pattern recognition in data mining, and image or signal processing.

**6.3 Reinforcement Learning**

In the machine learning paradigm known as reinforcement learning (RL), an agent gains the ability to make decisions in a particular order by interacting with its surroundings. Through taking actions, observing the results, and getting feedback in the form of rewards or penalties, the agent learns by making mistakes and learning from them. Learning the best course of action to maximize cumulative rewards over time is the aim of reinforcement learning for the agent.

Applications for reinforcement learning can be found in many different fields, including robotics, recommendation systems, autonomous systems, and game play (like AlphaGo). It works especially well in situations where a system must learn by interacting with its surroundings, making decisions one after the other to reach a long-term objective, and then modifying its plan in response to feedback.

1. **Deployment and Production**

**7.1 Deployment**

In the context of AI/ML, deployment refers to the process of making a machine learning model operational and accessible for usage in real-world applications. It involves implementing the trained model into a production environment where it can generate predictions or perform tasks based on new, unseen data.

Deployment is closely tied to the concept of production in AI/ML. When a model is in production, it means that it is live and actively serving predictions or making decisions based on real-time or batch data. In a production environment, the model is used to perform tasks, and its outputs are typically consumed by end-users or other systems.

The process of moving a model into production involves collaboration between data scientists, software engineers, DevOps teams, and domain experts. A smooth deployment into production ensures that the model operates effectively, delivering the intended value without disrupting the existing workflow. Production-level models have to meet various requirements such as reliability, speed, security, and compliance with industry standards.

The ultimate goal of deployment and production in AI/ML is to create a reliable, efficient, and usable system that leverages the benefits of machine learning to deliver value to users or organisations, solving real-world problems and improving decision-making processes.

**7.2 Scalability**

Scalability in the deployment and production of AI/ML systems refers to the ability of these systems to handle increased workloads, adapt to larger datasets, and accommodate higher traffic or computational demands while maintaining or improving performance. It's crucial to ensure that the deployed models and systems can efficiently manage increased volumes of data and user requests without sacrificing responsiveness or accuracy.

Scalability is essential in AI/ML deployment as it allows systems to handle increasing demands without degradation in performance, ensuring that models remain responsive, reliable, and accurate even under heavier workloads. Achieving scalability requires a combination of technological considerations, efficient design, and infrastructure choices, facilitating the seamless and efficient operation of AI/ML systems in various real-world applications.

To achieve scalability, organizations often leverage cloud services, containerization, and microservices architecture, allowing them to flexibly allocate resources and scale components of their AI/ML systems as required. Properly managed scalability enhances the efficiency and cost-effectiveness of AI/ML deployment and production, ensuring that the systems can grow and adapt to meet evolving demands.

**Objective**

The primary objective of an AI/ML project on credit card fraud detection is to develop and deploy a system that effectively identifies fraudulent credit card transactions. This project aims to improve security and trust in financial transactions, protect both consumers and financial institutions from financial losses, and ensure the smooth operation of the payment ecosystem. The key objectives of such a project include:

* ***Fraud Detection****:* Develop and train machine learning models to accurately detect fraudulent credit card transactions. These models should identify unauthorised or suspicious activities in real time.
* ***Real-time Processing****:* Implement real-time or near-real-time processing to detect and respond to fraudulent transactions as they occur, preventing potential losses.
* ***Data Privacy and Security****:* Ensure the protection of sensitive customer information during the fraud detection process, complying with data privacy regulations.
* ***Customer Experience****:* Prioritize a seamless and positive customer experience by minimizing disruptions caused by fraud detection mechanisms and providing clear communication in case of false positives.
* ***Cost Efficiency****:* Develop a cost-effective system that balances fraud detection effectiveness with operational costs.

Overall, the goal of a credit card fraud detection project is to strike a balance between detecting fraudulent activities with high accuracy, minimizing false positives, ensuring data security and compliance, and delivering a positive experience for both customers and financial institutions. The project's success is measured by its ability to protect the financial ecosystem and maintain trust in electronic payment systems.

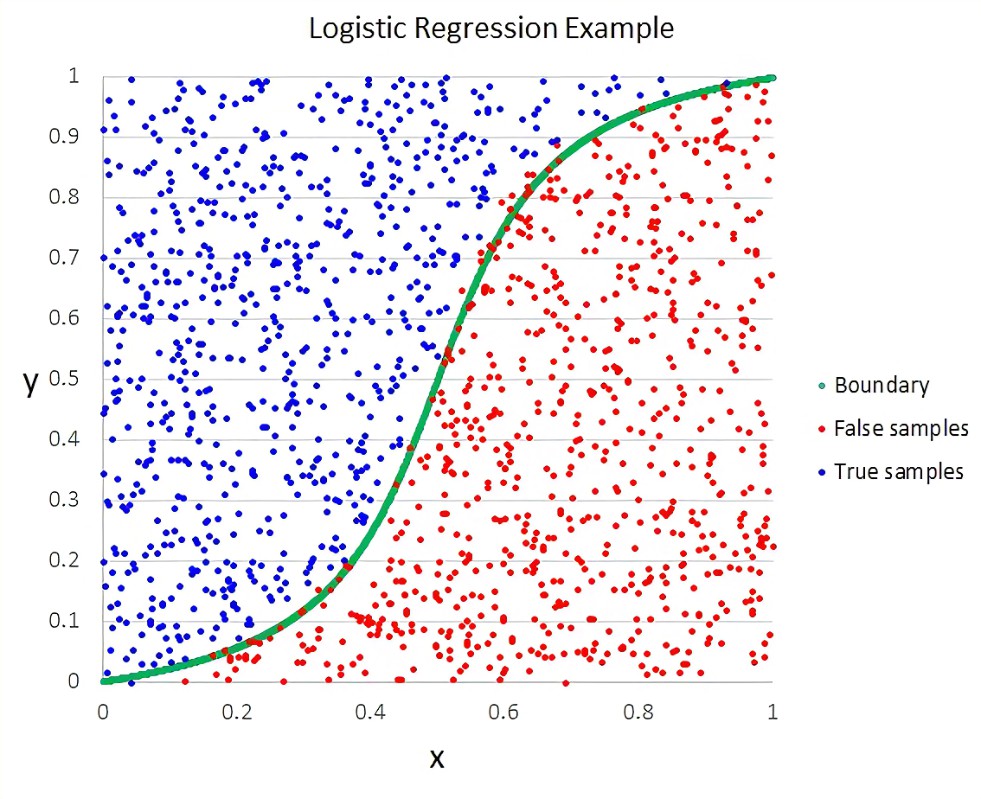
**Algorithm**

**1. Non-Linear Regression**

Non-linear regression is a statistical method used to model complex relationships between variables, allowing the estimation of non-linear patterns within data. It enables the understanding of intricate, non-linear patterns within data, beyond the capabilities of linear models. By employing diverse non-linear functions, such as polynomial, exponential, or sigmoidal functions, it captures complex correlations between variables. Non-linear regression is vital in machine learning for tasks where relationships are not linear, allowing models to better represent phenomena like growth rates, saturation effects, or irregular patterns, enhancing predictions in various domains, including image and speech recognition, financial forecasting, and scientific research.

**Logistic Regression**

Logistic regression is a fundamental and widely used statistical method in machine learning for binary classification tasks. Despite its name, logistic regression is a classification algorithm, not a regression technique used for continuous output prediction. It is used to predict the probability of occurrence of a binary event based on one or more predictor variables. This algorithm is particularly valuable in fields such as finance, healthcare, marketing, and more.

****

**Understanding Logistic Regression:**

**Basic Principle:**

Logistic regression predicts the probability of a binary outcome by applying the logistic function (also called the sigmoid function) to a linear combination of input features.

It models the relationship between the independent variables (features) and the probability of a certain class or outcome occurring.

**Sigmoid Function:**

The sigmoid function transforms any real-valued number into a range between 0 and 1. It's defined as *σ*(*z*)= 1/(1+*e*^-z) ​where z represents the linear combination of input features.

sigmoid function converts the continuous variable data into the probability i.e. between 0 and 1.

* tends towards 1 as  
* tends towards 0 as 
* is always bounded between 0 and 1

**Key Features:**

* It's a parametric model that uses coefficients to learn and predict outcomes.
* It's suited for binary classification problems but can be extended to multi-class classification using one-vs-rest or one-vs-one strategies.
* It's sensitive to outliers and assumes that the relationship between the input features and the output is linear.

**Application of Logistic Regression:**

* ***Binary Classification:*** Predicting outcomes with two classes, such as spam vs. non-spam emails, disease vs. no disease, etc. It's used in risk assessment, fraud detection, and credit scoring.
* ***Probabilistic Interpretation:*** Logistic regression provides probabilities rather than direct classifications, making it useful in cases where understanding the certainty of predictions is essential.
* ***Evaluating Model Performance:*** Performance metrics like accuracy, precision, recall, F1 score, and ROC-AUC are often used to evaluate a logistic regression model.

**Logistic Regression Process:**

1. ***Data Preprocessing:*** Handling missing values, encoding categorical variables, and scaling features for better model performance.
2. ***Model Training:*** Optimizing the model's coefficients by iteratively updating them based on the error between predictions and actual outcomes.
3. ***Regularization:*** Regularization techniques like L1 (Lasso) and L2 (Ridge) regularization are often applied to prevent overfitting.
4. ***Model Evaluation:*** Using various metrics and techniques, like confusion matrices, precision-recall curves, or ROC curves, to assess the model's performance.

**Advantages of Logistic Regression:**

* ***Interpretability:*** Provides clear and interpretable results, allowing the understanding of how each input feature contributes to the prediction.
* ***Efficiency and Speed:*** It's computationally efficient and less complex than other more advanced algorithms.
* ***Ease of Implementation:*** Relatively easy to implement and doesn’t require extensive computational resources or sophisticated tuning.

**Limitations of Logistic Regression:**

* ***Linearity Assumption:*** It assumes a linear relationship between input features and the log odds of the outcome, restricting its ability to model non-linear relationships.
* ***Vulnerability to Outliers:*** Sensitive to outliers that can skew the coefficients and the overall model.
* ***Assumption of Independence:*** Assumes that the input features are independent, which might not be true in real-world data.

In conclusion, logistic regression is a foundational method in binary classification tasks, balancing simplicity, interpretability, and effectiveness. While it has its limitations, its practicality and ease of implementation make it an invaluable tool in various applications within the field of machine learning and statistics.

**Random Forest Regression:**

Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

Random Forest is an ensemble technique capable of performing regression and classification tasks using multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as bagging. The basic idea behind this is combining multiple decision trees to determine the final output rather than relying on individual decision trees.

Random Forest has multiple decision trees as base learning models. We randomly perform row and feature sampling from the dataset, forming sample datasets for every model. This part is called Bootstrap.

Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output. Therefore, below are two assumptions for a better Random Forest classifier:

* There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.
* The predictions from each tree must have very low correlations.

Every decision tree has high variance, but when we combine all of them in parallel, the resultant variance is low as each decision tree gets perfectly trained on that particular sample data. Hence, the output depends not on one but multiple decision trees. In the case of a classification problem, the final output is taken using the majority voting classifier. In the case of a regression problem, the final output is the mean of all the outputs. This part is called Aggregation. Random forest is an ensemble of decision trees. This is to say that many trees, constructed in a certain “random” way form a Random Forest.

* Each tree is created from a different sample of rows and at each node, a different sample of features is selected for splitting.
* Each of the trees makes its own individual prediction.
* These predictions are then averaged to produce a single result.

The averaging makes a Random Forest better than a single Decision Tree hence improves its accuracy and reduces overfitting.

A prediction from the Random Forest Regressor is an average of the predictions produced by the trees in the forest. Advantages of Random Forest Regression:

1. It takes less training time as compared to other algorithms.
2. It predicts output with high accuracy, even for the large dataset it runs efficiently.
3. It can also maintain accuracy when a large proportion of data is missing.
4. It is capable of performing both Classification and Regression tasks.
5. It is capable of handling large dataset ts with high dimensionality.
6. It enhances the accuracy of the model and prevents the overfitting issue.

Based on the accuracy metrics provided for the different models, it seems that all three models, logistic regression, random forest, and decision tree, perform exceptionally well with very high accuracy scores on both the training and test data. However, there are several factors to consider when choosing the best alternative for your website:

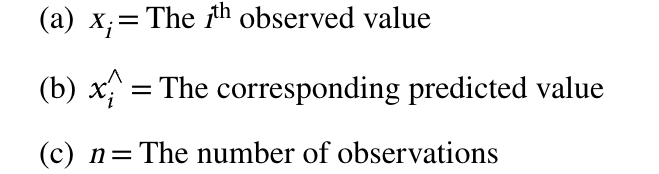
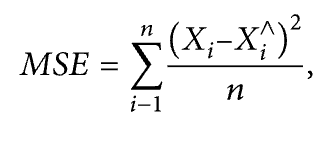
* ***Performance Consistency:*** While all models have high accuracy, you should evaluate the consistency of performance across different datasets and potential variations in the input data.
* ***Model Complexity:*** Consider the complexity of the models and their interpretability. Logistic regression and decision trees are generally easier to interpret compared to random forests.
* ***Computational Efficiency:*** Evaluate the computational resources required for each model, especially if you expect high traffic volume or real-time predictions on your website.
* ***Scalability and Maintenance:*** Consider each model's scalability and ease of maintenance. Some models may require more frequent updates and retraining to maintain optimal performance over time.

Based on these factors, logistic regression might be the preferred option due to its simplicity, interpretability, and good performance in this case. However, it's essential to consider your website's specific requirements and constraints and choose the model that best aligns with your goals and constraints.

**Metrics Used**

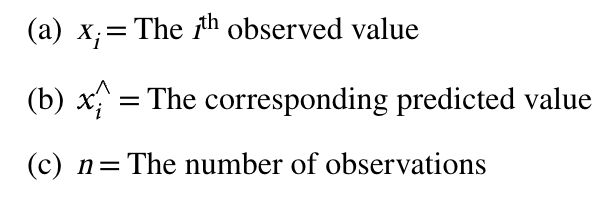
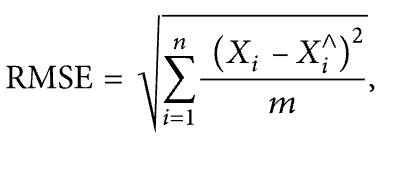
* **Mean Squared Error (MSE):**

MSE is a parameter that measures how closely a fitted line resembles a set of data points. The lower the value, the closer it is to the line, and hence the better. If the MSE value = 0, the model is perfect. It is shown in equation



* **Root Mean Squared Error (RMSE):**

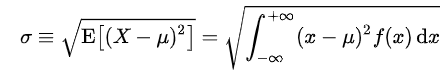
MAE evaluates the absolute distance of the observations to the predictions on the regression line. It is shown in the equation



* **Standard Deviation (SD):**

SD is a measure of the amount of variation of a set of values. A low standard deviation indicates that the values tend to be close to the mean of the set, while a high standard deviation indicates that the values are spread out over a wider range.

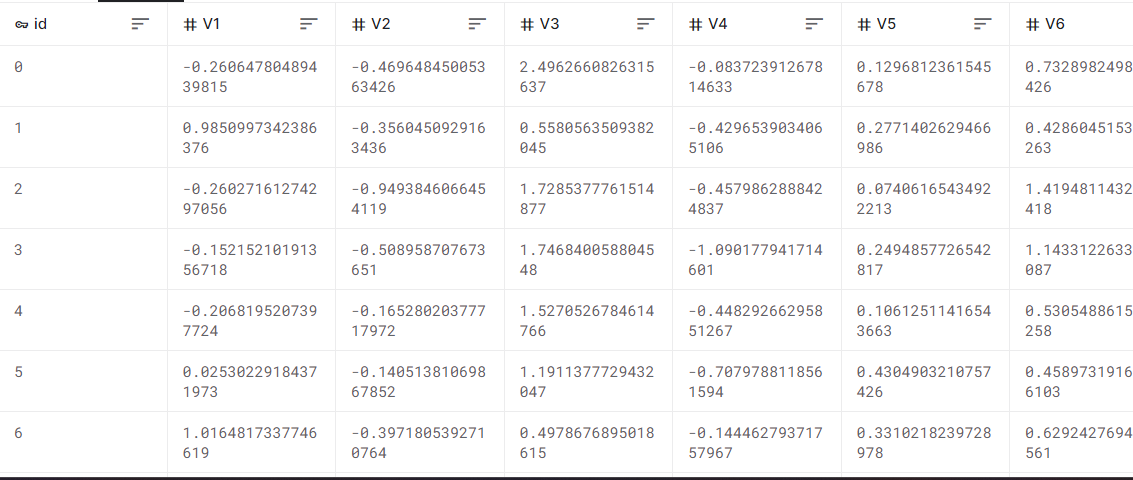
The standard deviation *σ* of *X* is defined as:



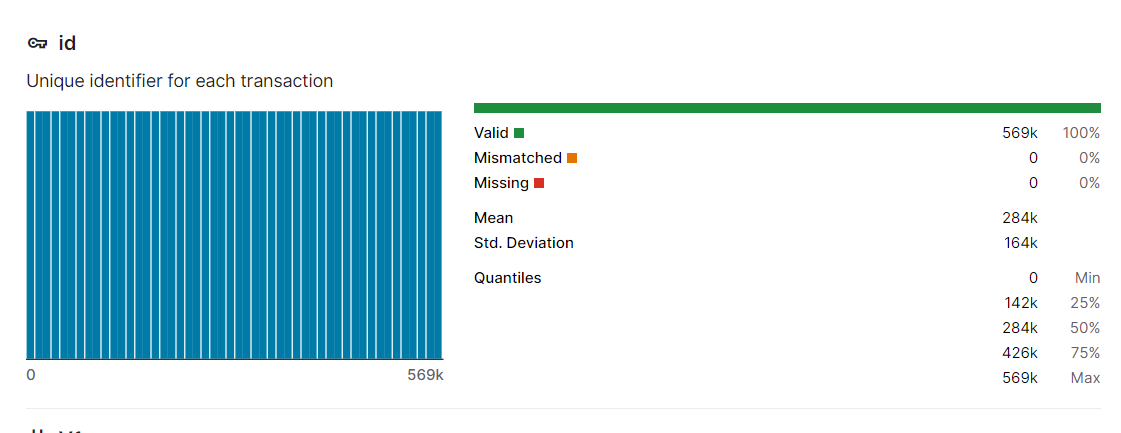
*Where μ* is the average of random variable *X* with density f(*x*)

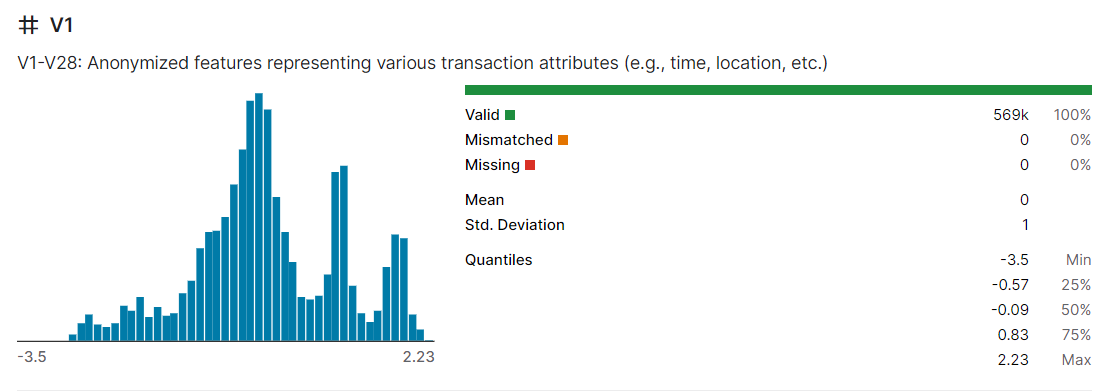
**Results And Discussion**

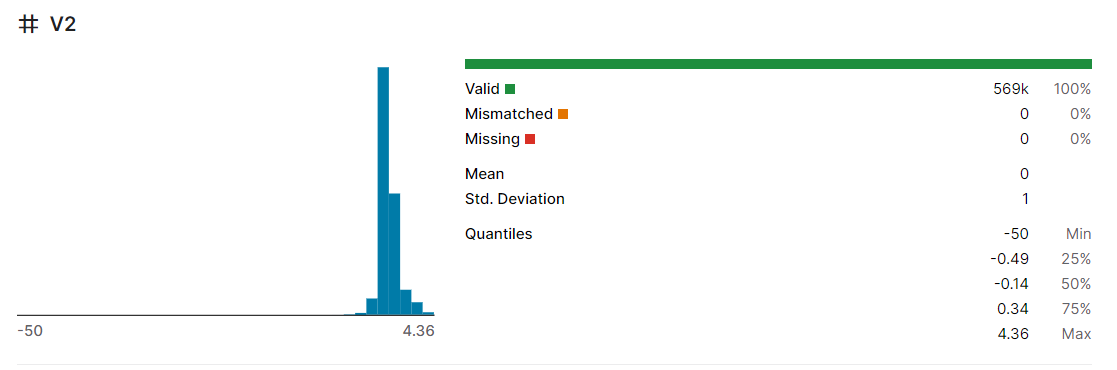
* **Data in Compact Form**

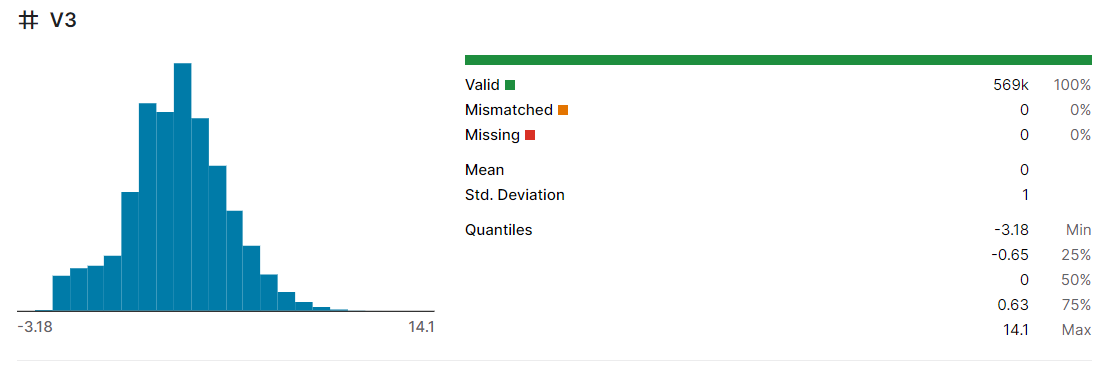
****

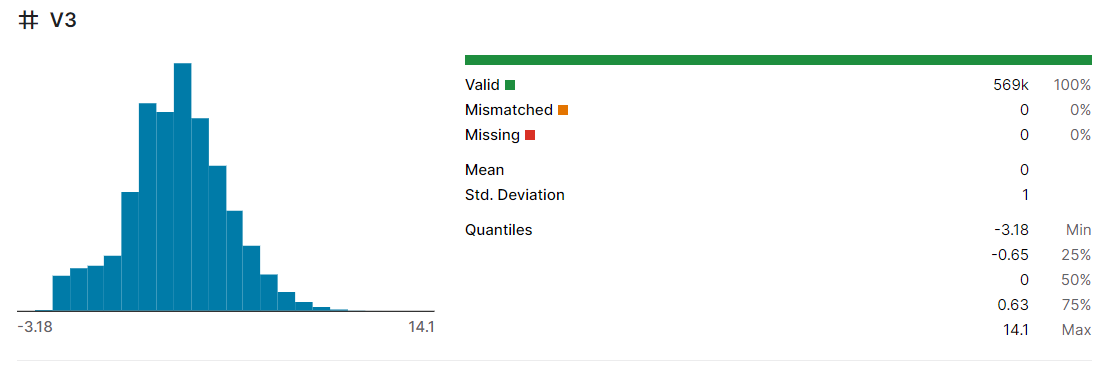
* **Data in Column Form**

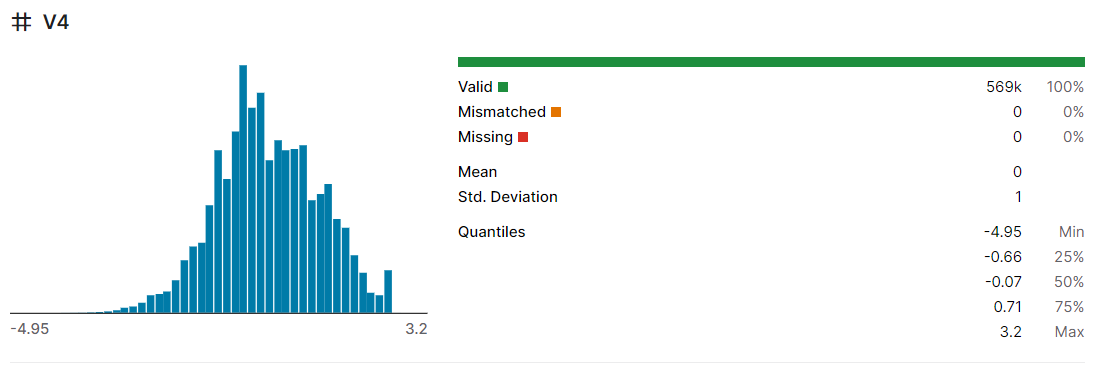
****

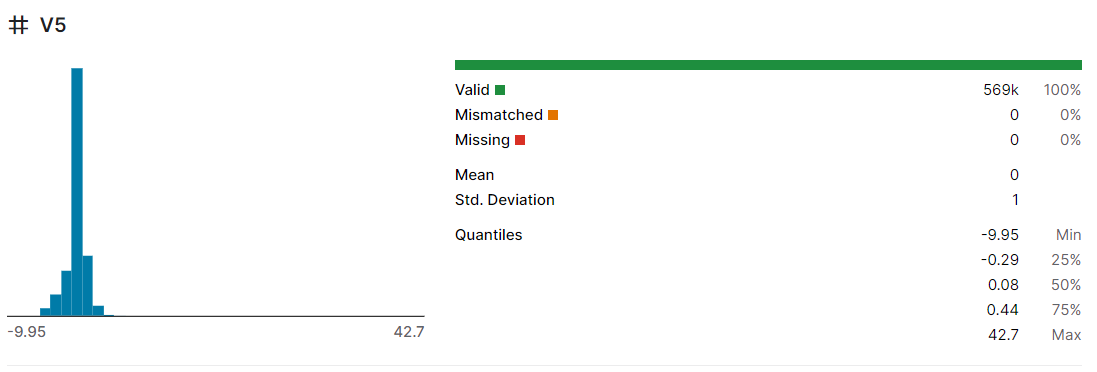
****

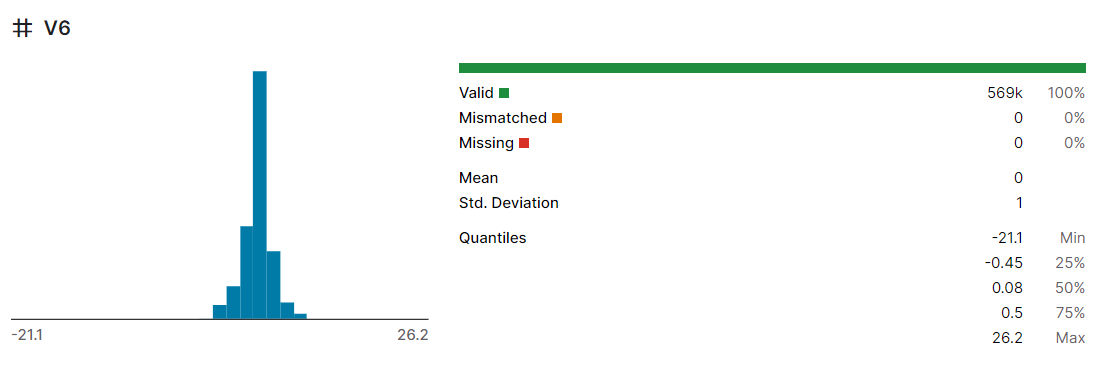
****

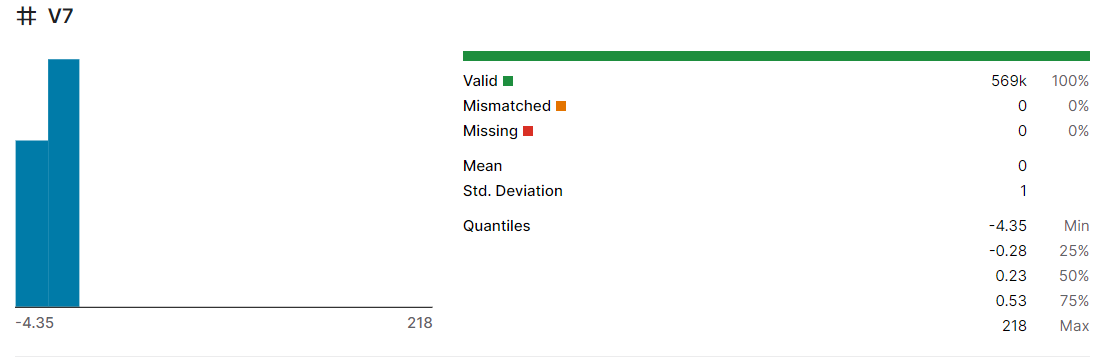
****

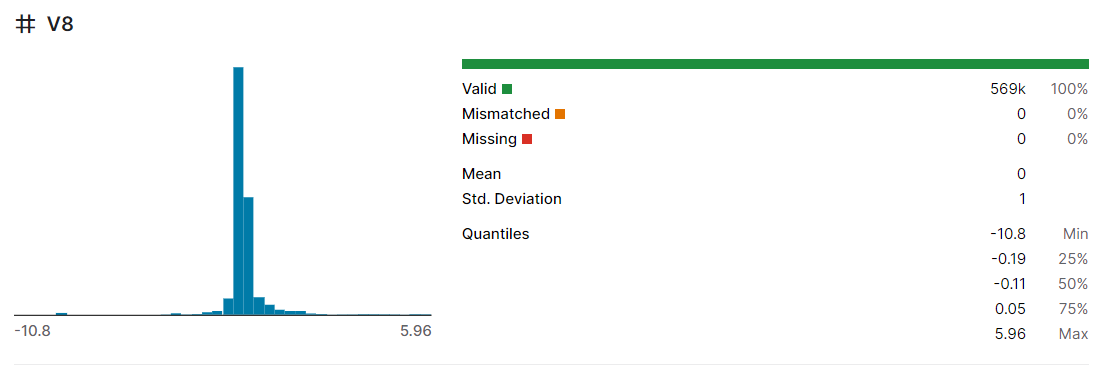
****

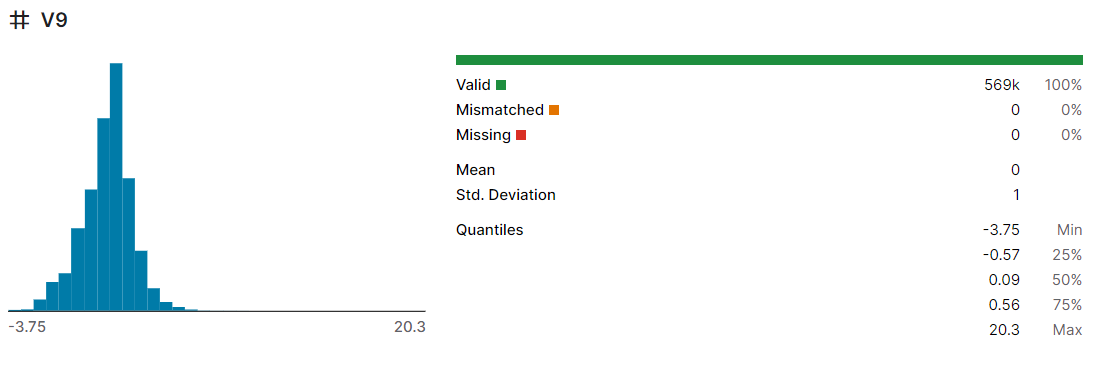
****

****

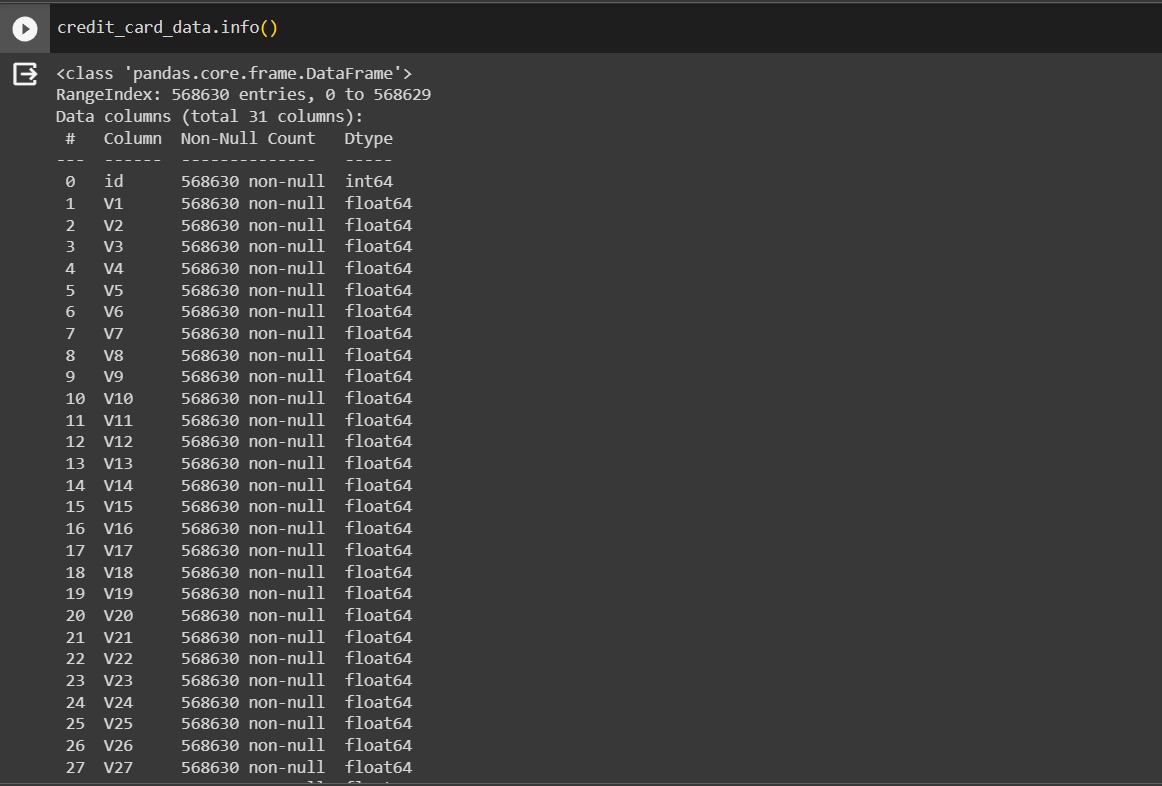
****

****

****

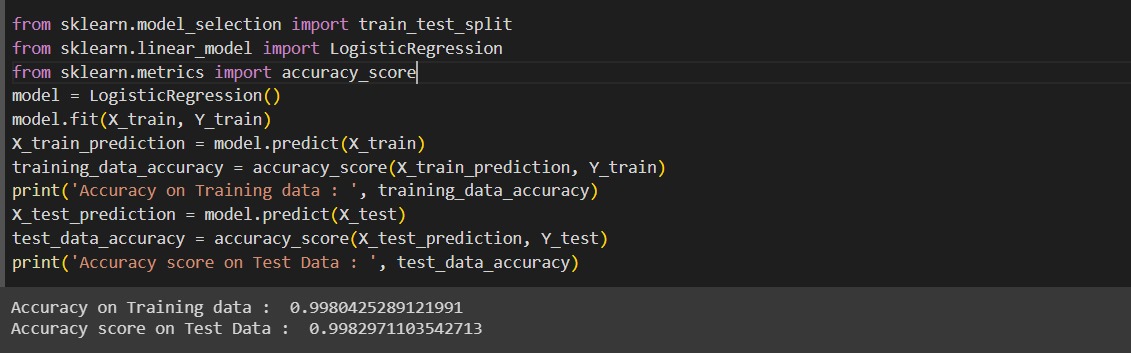
****

**Data Processing**



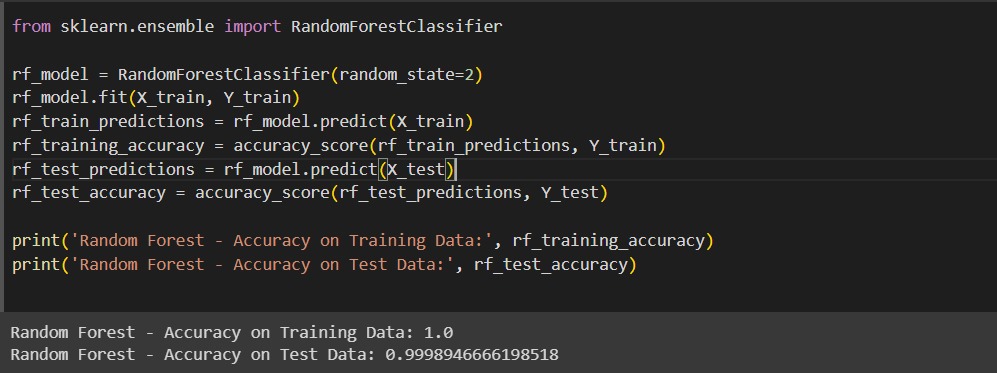
**Logistic Regression:**

**Code:**



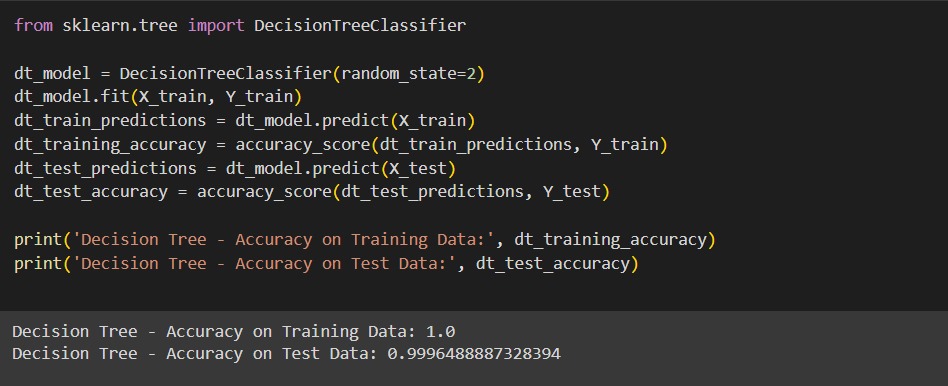
**Random Forest:**

**Code:**



**Decision Tree:**

**Code:**



**Accuracy Comparison of All Models:**

**Code:**

models = ['Logistic Regression', 'Random Forest', 'Decision Tree']

train\_accuracies = [lr\_training\_data\_accuracy, rf\_training\_accuracy,dt\_training\_accuracy]

test\_accuracies = [lr\_test\_data\_accuracy, rf\_test\_accuracy, dt\_test\_accuracy]

x = np.arange(len(models))

width = 0.35

fig, ax = plt.subplots()

rects1 = ax.bar(x - width/2, train\_accuracies, width, label='Training Accuracy', color='b')

rects2 = ax.bar(x + width/2, test\_accuracies, width, label='Testing Accuracy', color='g')

ax.set\_ylabel('Accuracy')

ax.set\_title('Model Comparison - Training and Testing Accuracy')

ax.set\_xticks(x)

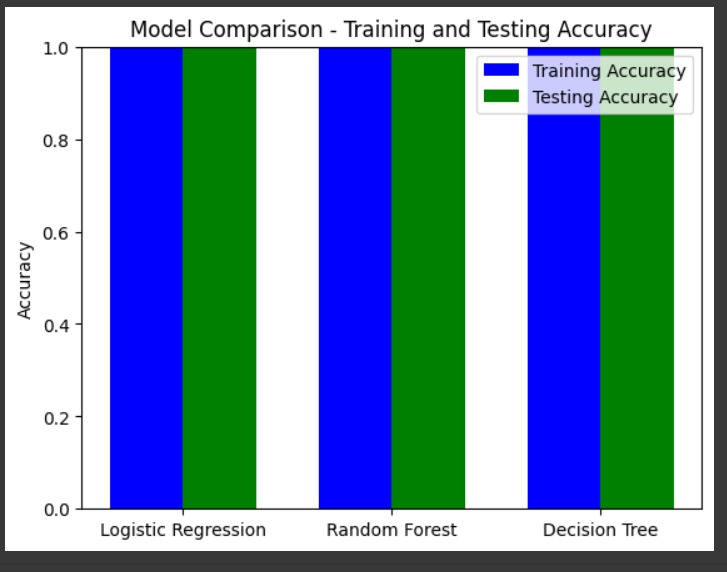
ax.set\_xticklabels(models)

ax.legend()

plt.ylim(0.0, 1.0)

plt.show()

**Output:**

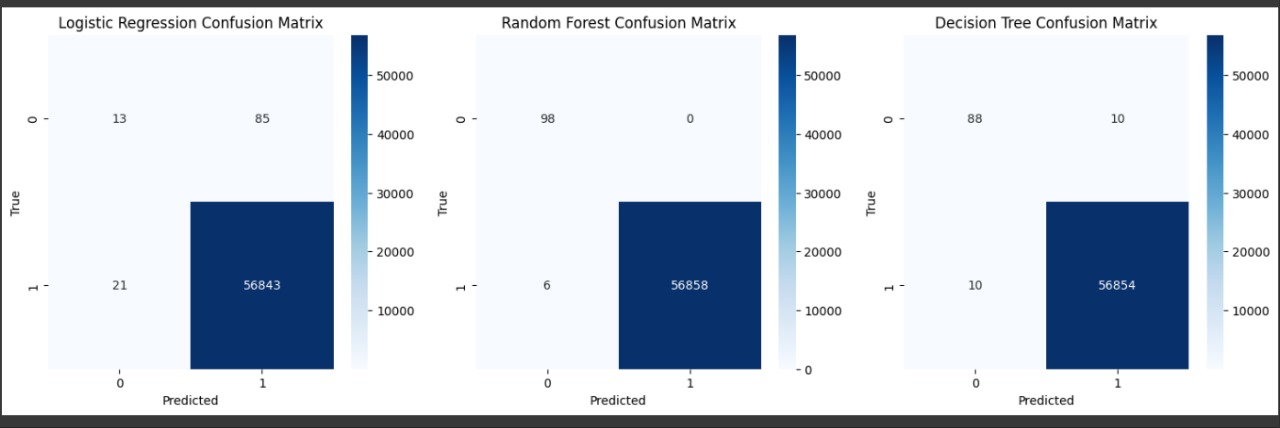


**Confusion Matrices of all Models:**

**Code:**



**Output:**



**Deployment**

**Brief Introduction to a Web App**

A web application is an application software that runs on a web server, unlike computer-based software programs that are stored locally on the Operating System of the device. Web applications are accessed by the user through a web browser with an active internet connection. These applications are programmed using a client–server modeled structure—the user is provided services through an off-site server that is hosted by a third-party. The third party whose services we are using is Render. Render is a great place to launch your apps upto a limited size. We could not do this project without the support of Render who gave us the opportunity to host our web app.

**Behind the scenes**

In case of any web app, everyone sees the front page of it, we will discuss that in the next section, in this section, we will be discussing what is happening in behind.

We are using various files to host this web app, the directory structure can be seen as below:

App **app.py** is the main application file which will run our app, it will be the first file which will be called and executed, after that, this will render the templates, css styles and will also take input given by the user.

**Requirements:** requirements.txt file will contain all the requirements we will be going to use for our work, i.e. model prediction, reading files, reading model files etc.

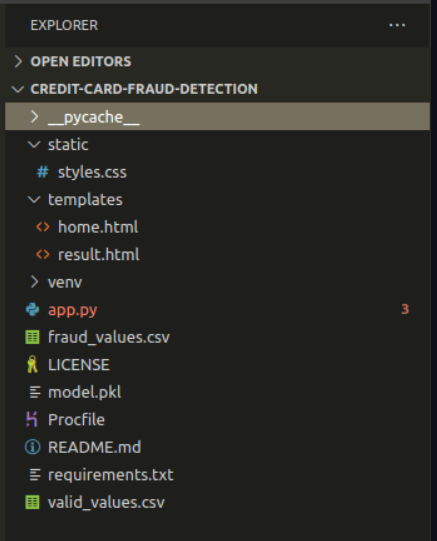
**Templates:** templates directory contains the html pages we are going to render upon.

**Static:** static contains the css styles which are used for the better formatting of the web app.

**Model:** model.pkl contains the model we got as a result of training which we discussed in the previous section.

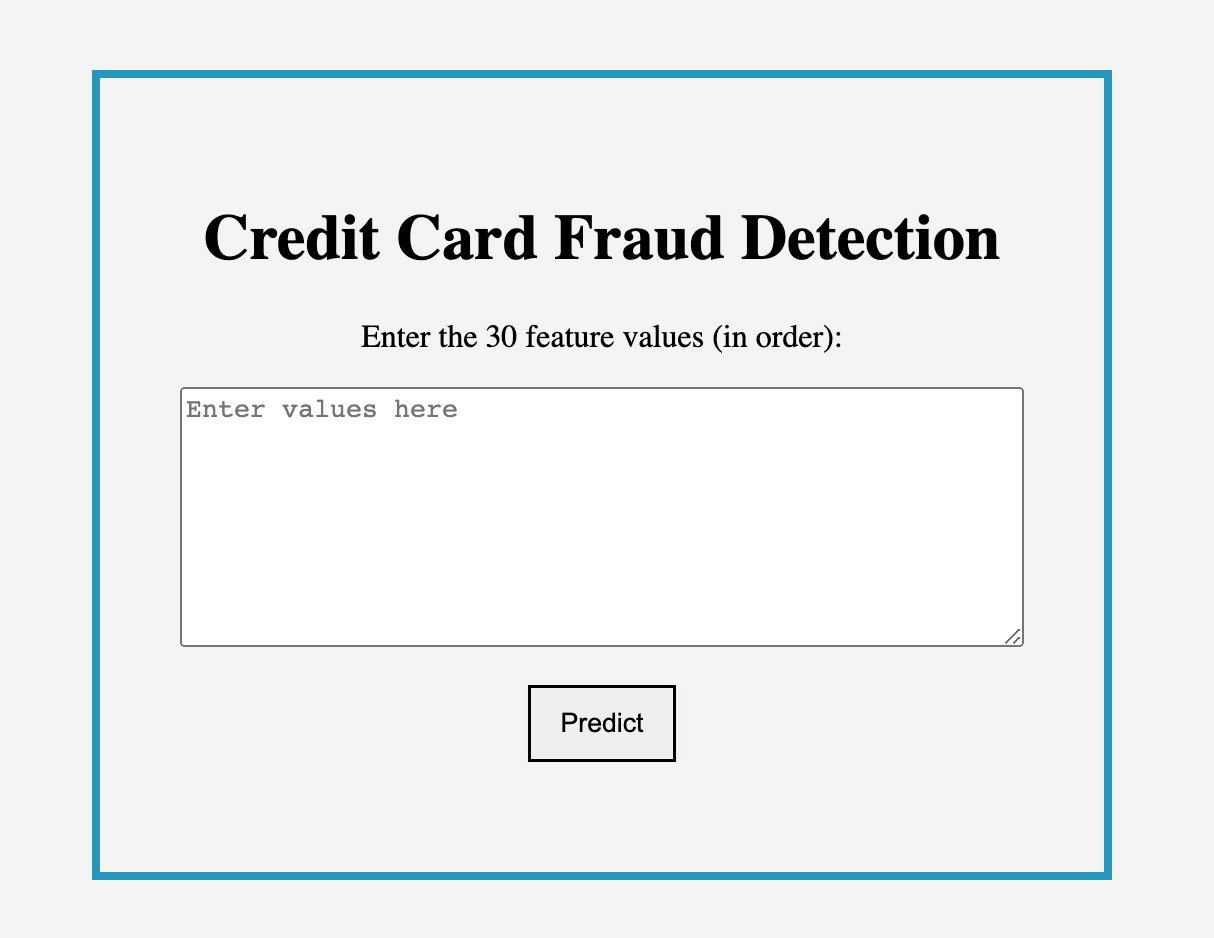
**Test data fraud\_values.csv and valid\_values.csv:** are the two files which contains the test data regarding the testing of the web app for our users.

**Procfile:** Procfile is the one file which is most important and mandatory if we want to host our app on Heroku, procfile contains the details about the type of app and which file we want to run while starting the app(in our case, it is app.py).

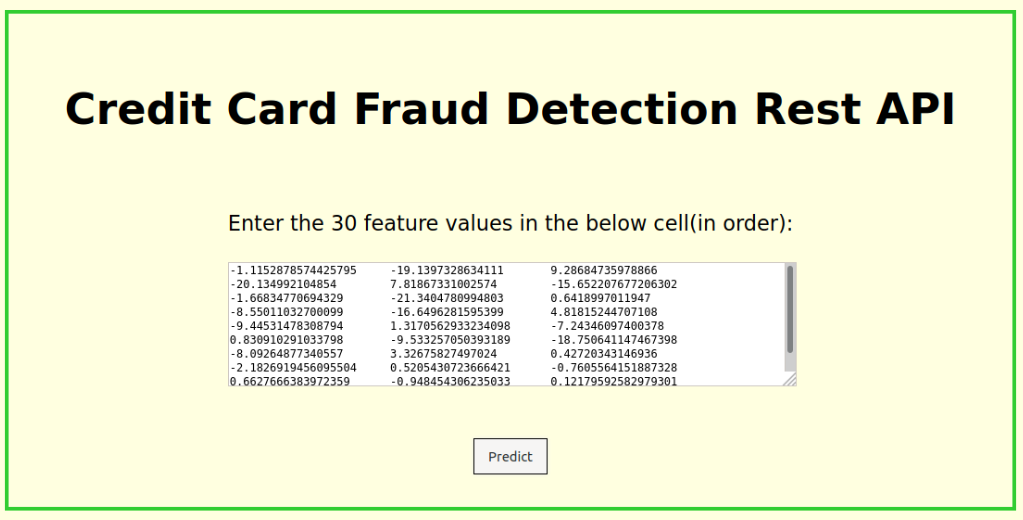


**How to Predict Using Web App**

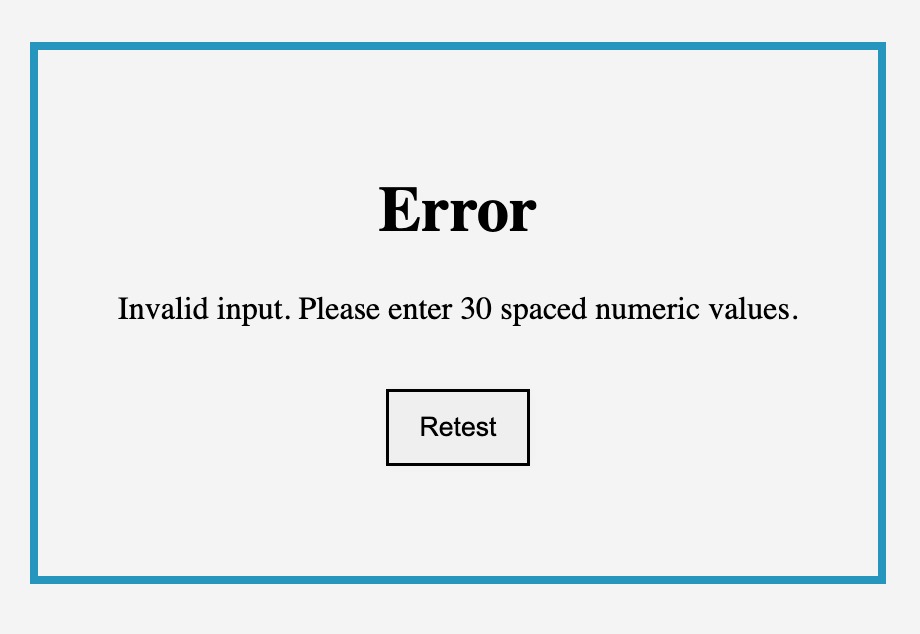
Home The home page of the web app will look something like this:

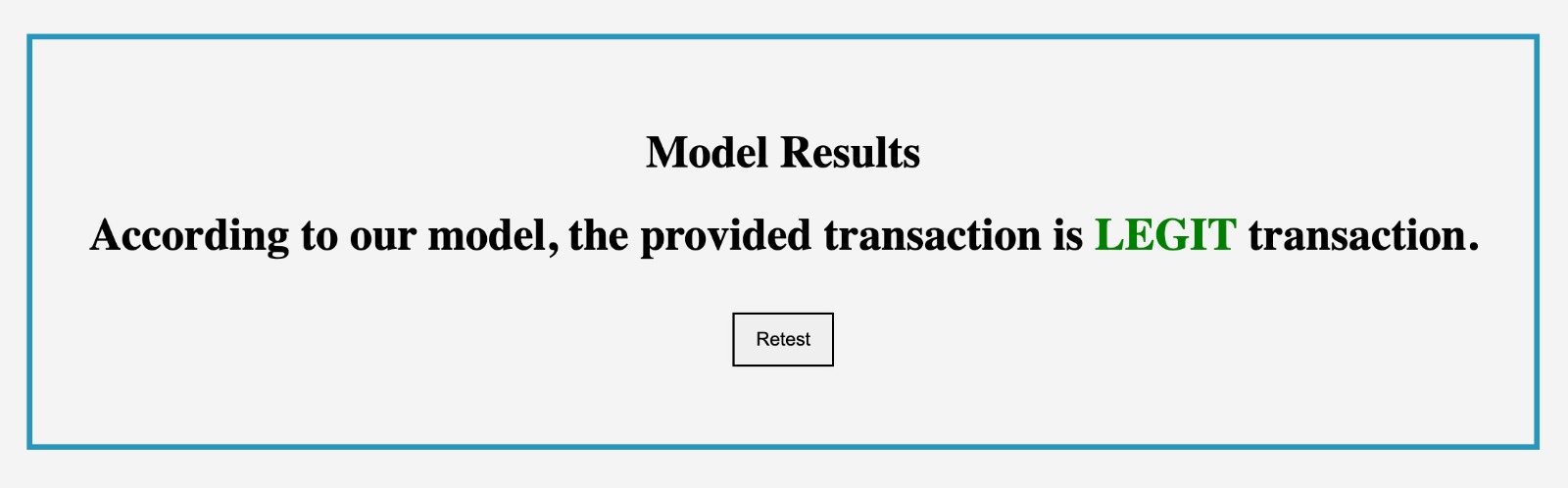


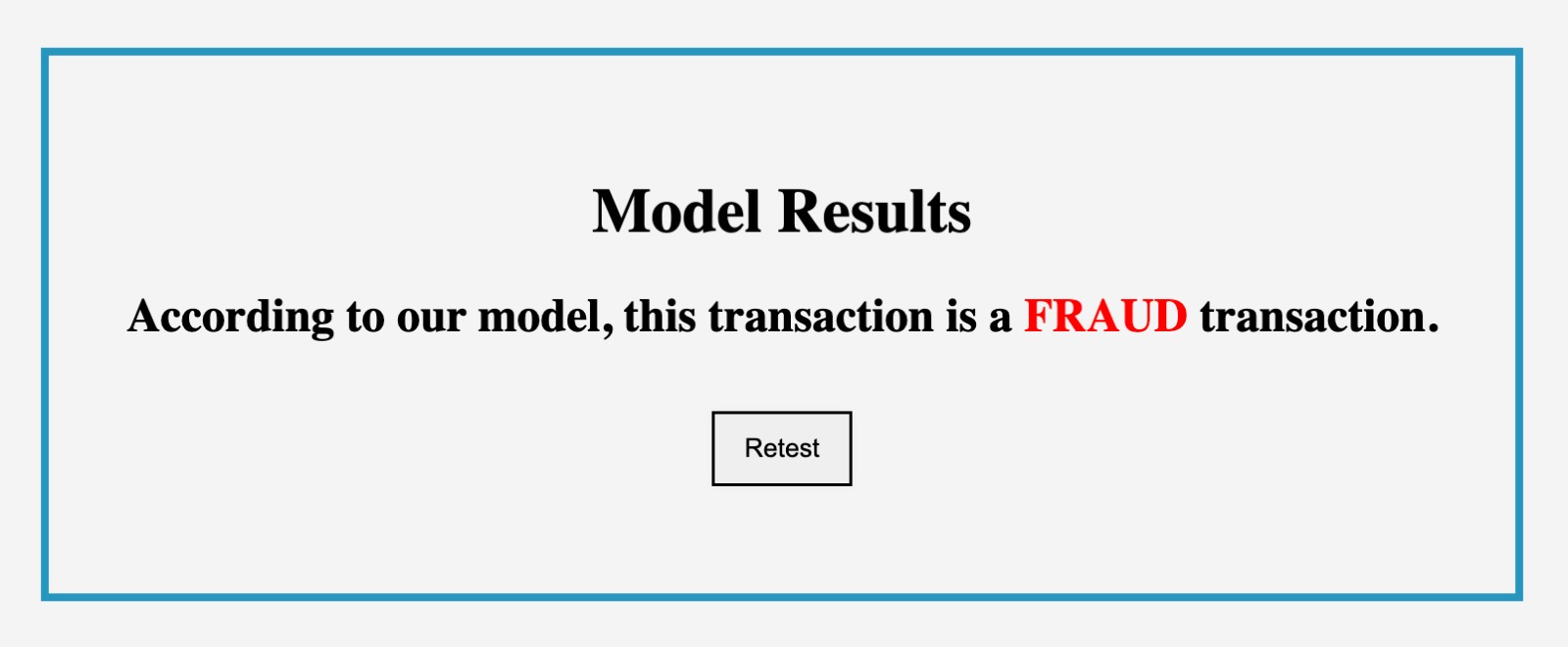
What we need to do is just copy the data entry which we want to predict into the box, this box will take 30 float/integer values as input, with at least one whitespace dividing them.



**Outputs:**





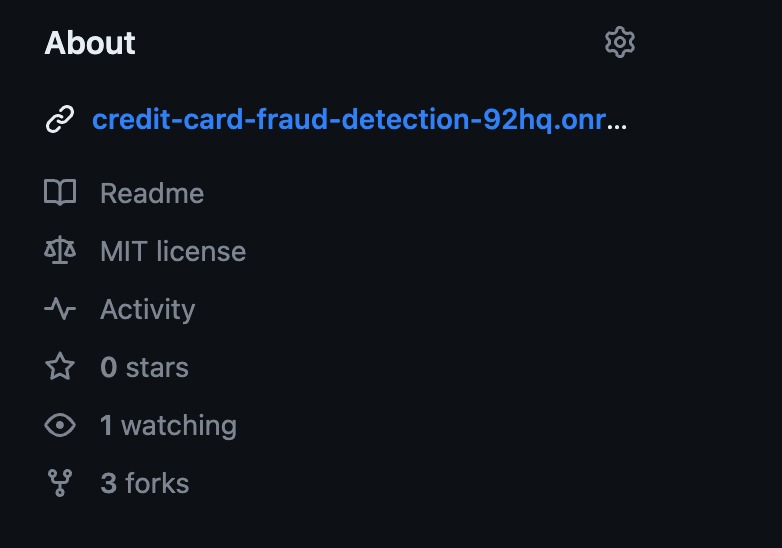


**Website Link**

**Go to Github Repository**

<https://github.com/singh-anushka/Credit-Card-Fraud-Detection>

**Under about section**



**Conclusion**

In conclusion, our credit card fraud detection project, which employed the logistic regression algorithm, has proven to be a valuable and effective approach to safeguarding financial transactions from fraudulent activities. Through the course of our research and analysis, we have gained valuable insights and achieved several key outcomes:

1. ***Effective Fraud Detection:*** Our implementation of logistic regression exhibited strong capabilities in distinguishing between legitimate and fraudulent transactions. The model's ability to predict and classify fraud cases with a high degree of accuracy provides a robust foundation for fraud prevention.
2. ***Scalability and Efficiency:*** Logistic regression is a computationally efficient algorithm, making it suitable for real-time or large-scale credit card transaction monitoring. Its scalability and speed ensure timely detection and response to fraudulent activities.
3. ***Interpretability:*** Logistic regression offers transparency in the form of interpretable coefficients, which allows for a clear understanding of the factors that influence fraud detection. This transparency can be crucial in addressing regulatory compliance and improving decision-making processes.
4. ***Customization and Adaptability:*** The logistic regression model can be easily fine-tuned to adapt to changing fraud patterns and evolving tactics used by fraudsters. This adaptability is a significant advantage in the ongoing battle against fraud.
5. ***Reduced False Positives:***Through parameter tuning and feature selection, we were able to minimize the rate of false positives, ensuring that legitimate transactions are not mistakenly flagged as fraudulent. This reduction in false alarms is essential to maintain a positive user experience.
6. ***Continuous Improvement:*** The effectiveness of our logistic regression model can be further enhanced by incorporating additional data sources, utilizing more advanced feature engineering techniques, and employing ensemble methods to achieve even higher accuracy in fraud detection.

However, it is important to note that no single model or algorithm can provide complete immunity against credit card fraud. Fraudsters continuously evolve their tactics, requiring constant vigilance and adaptation. Our project is a stepping stone in the ongoing effort to combat fraud, and it underscores the importance of combining various machine learning techniques, along with expert oversight, to create comprehensive and robust fraud detection systems.

As we move forward, it is essential to collaborate with financial institutions, regulatory bodies, and the broader data science community to share knowledge and best practices, fostering a united front against fraudulent activities in the financial sector. Our project is a testament to the potential of logistic regression in credit card fraud detection, and we hope it inspires further research and innovation in this critical field.

In the end, our commitment to enhancing security in the financial industry remains unwavering. We believe our work will contribute to a safer and more secure environment for all credit card users.

**References**

1. https://jovian.com/biraj/deploying-a-machine-learning-model

2. https://www.kaggle.com/datasets/kartik2112/fraud-detection

3. https://css-tricks.com/snippets/css/a-guide-to-flexbox/

4. https://www.youtube.com/watch?v=NCgjcHLFNDg