**CREDIT CARD APPROVAL PREDICTION**

**TEAM MEMBERS**

* Jeevan Sravanthi Paidi
* Charitha Sai Kaza
* Likhitha Reddy Patlolla
* Nikhila Putta
* Srikar Bandla

**ABSTRACT**

•The goal of this project is building a machine learning model that can predict credit card approval or not. The model will be trained on a large dataset of credit card applications and will use various features such as income, credit score, employment status, and loan history to predict the likelihood of approval.

•The key factor for these predictions is credit score which is a common risk management tool in the financial industry. Credit scores can quantify the magnitude of risk objectively.

•The data is trained using different machine learning algorithms then the model will be evaluated on various metrics such as accuracy, p-tests, and will be analyzed using appropriate statistical methods.

**DATA SET DESCRIPTION**

●The credit\_record dataset has 1048575 records with 3 columns.

●The application\_record.csv dataset has 438557 records with 18 columns.

●Income is negatively correlated with the number of days employed.

●The mean income of approved credit card applicants is higher than the mean income of rejected applicants.

●The mean income of approved credit card applicants is higher than the mean income of rejected applicants.

​​**POTENTIAL STATISTICS TESTS**

•One sample T-test can be used to compare the means of two groups, such as the mean income of male and female applicants.

•Fisher's exact test is performed in order to identify any non - random relationships among two category variables.

•Two samples unpaired t-test it compares the mean of two independent groups.

•Gaussian Distribution is a continuous probability distribution for a random variable with a real value.

**DATA SPECIFICATION**

* In the combined dataset we have 20 features where feature ID does not impact to predict output, so it was dropped.
* Finally, we have 19 features with 777715 observations.
* There are 308790 duplicate observations.
* We have 14 categorical features and 5 numerical features.
* Most features are symmetrically distributed except they are skew symmetrically distributed.

**LIST OF FEATURES**

1.CODE\_GENDER: Describes gender of customers.

2.FLAG\_OWN\_CAR: It says users have a car or not.

3.FLAG\_OWN\_REALTY: It says the users have a property or not.

4.CNT\_CHILDREN: It describes the number of children.

5.AMT\_INCOME\_TOTAL: It says the total annual income.

6.NAME\_INCOME\_TYPE: It describes the type of income source.

7.NAME\_EDUCATION\_TYPE: It describes the level of education.

8.NAME\_FAMILY\_STATUS: It describes whether the customer is married or not.

9.NAME\_HOUSING\_TYPE: It says the type of house they are living i.e house/apartment

10.DAYS\_BIRTH: It describes the number of days from birth.

11. DAYS\_EMPLOYED: It says number of days they are employed

12. FLAG\_MOBIL: It says whether they had a mobile or not.

13. FLAG\_WORK\_PHONE: It says whether they had a work phone or not.

14. FLAG\_PHONE: It says whether they had a telephone or not.

15. FLAG\_EMAIL: It says whether they had email or not.

16. OCCUPATION\_TYPE: It describes the type of work they are doing.

17. CNT\_FAM\_MEMBERS: It tells the number of family members.

18. MONTHS\_BALANCE: It describes the monthly balance.

19. STATUS: It has 5 different categories which have been classified based on the past due days. We have mapped those 5 categories into 2 classes i.e 0 or 1

**DESIGN AND MILESTONES**



**Data Collection** : we have collected the dataset from Kaggle repository

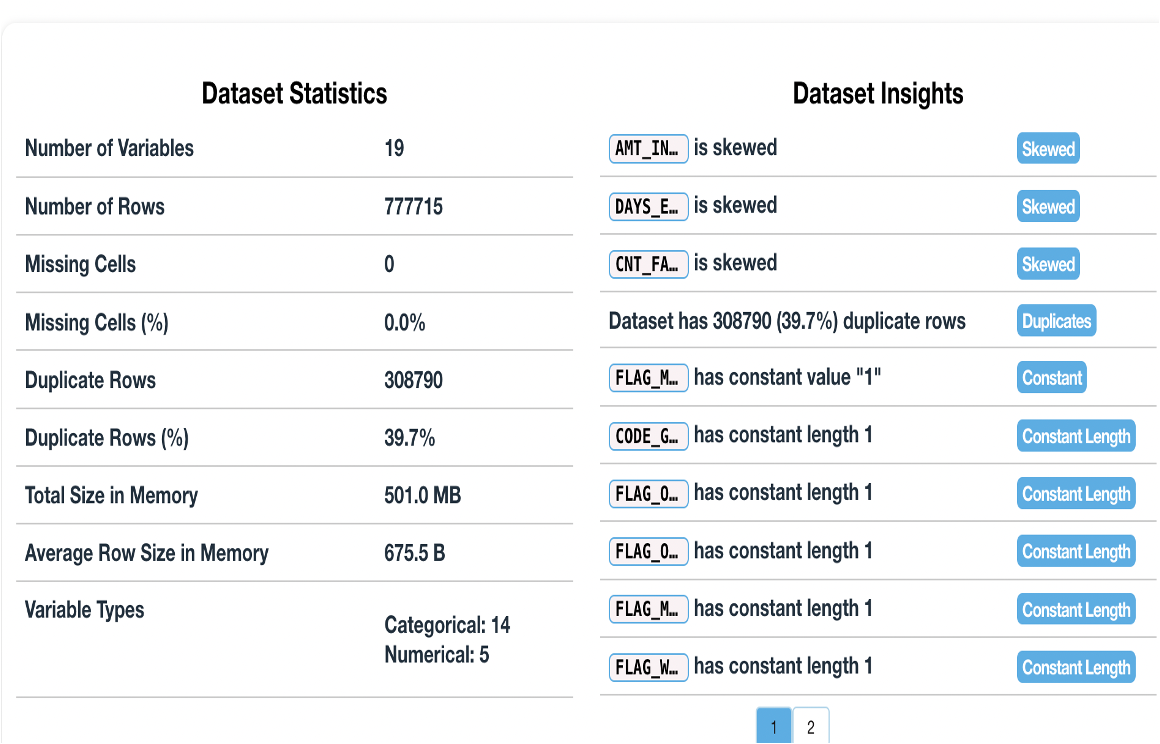
**Data Analysis** : This is a key milestone in this project, where we have used different data analysis libraries and frameworks.

• create\_report is to generate profile reports from a pandas DataFrame.

• create\_report utilizes the functionalities and formats the plots from dataprep.

• pandas\_profiling is to provide a one-line Exploratory Data Analysis (EDA) experience in a consistent and fast solution.

Figure shows the overview of the data that is drawn form create\_report library.



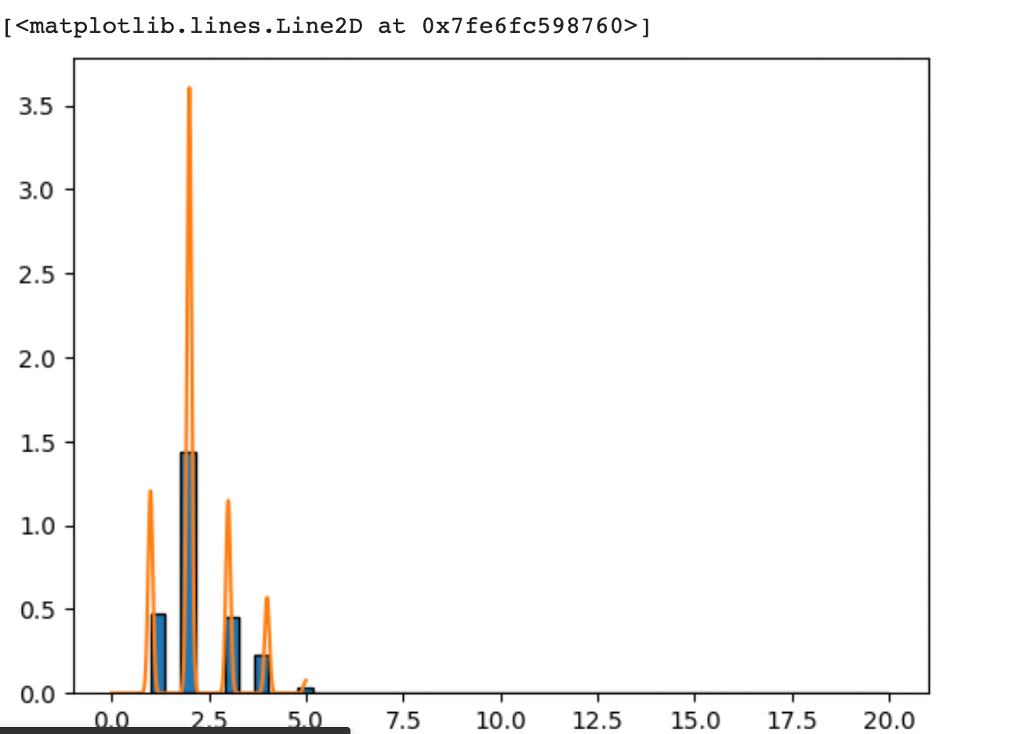
•During the data analysis phase we came across the potential statistical methods that we can perform on the data.

•Those are one sample t test, two sample unpaired t test, fisher's Exact test on features hypothesis and we calculated the corresponding p-value, then we get all p-values less than 0.5, hence we rejected the null hypothesis.

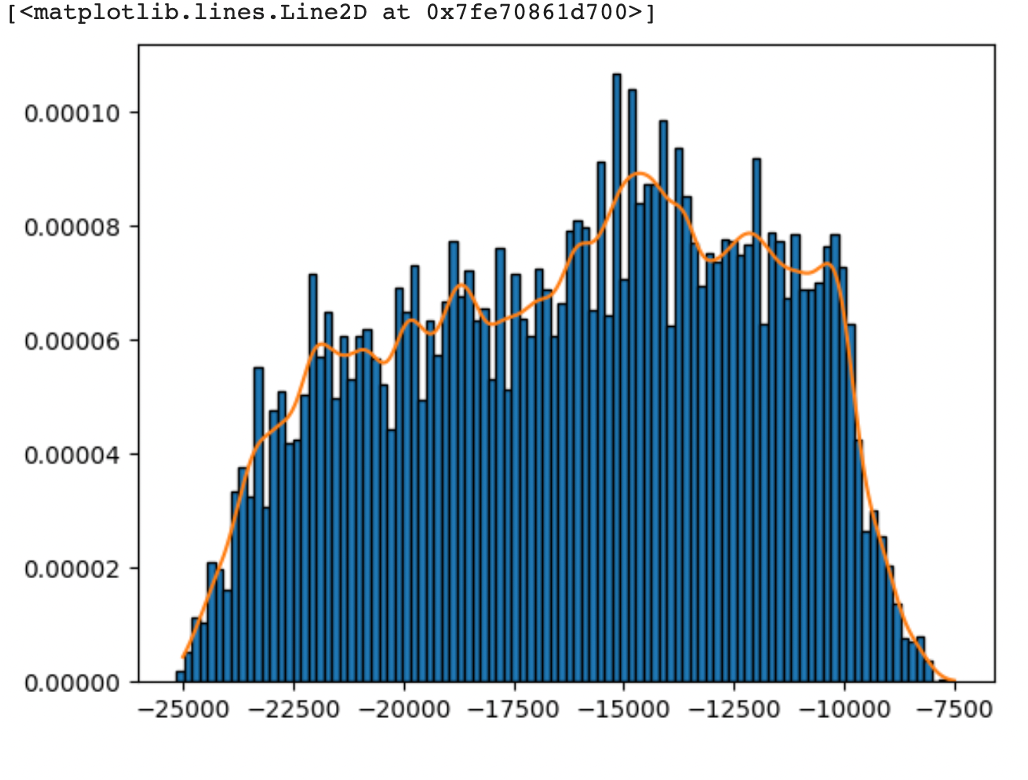
•And also, we tested the distribution of our data, most of them are normally distributed except some are skewed in its direction.

• We also plot the gaussian distributions for some features using the histograms. We showed that below.

* The graph represents the gaussian distribution of the number of family members.



* The graph represents the gaussian distribution of the number of days from birth.



**Data Preprocessing**

•This is also a major milestone, which improved the data quality, readability to make it consistent to run machine learning models.

•We have used the following libraries and tools for preprocessing.

•klib for basic data cleaning like removing duplicate rows, dropping low correlated features.

•Mapped the multiclass target feature to binary class i.e 0 or 1.

•Label encoding on object or categorical type features. Which will encode the labels to values.

•Sliced dependent and independent variable and split train and test data in 80:20 ratio.

•Oversampling the minority classes using SMOTE technique.

•Standardize each feature by removing the mean and scaling to unit variance using StandardScaler( ).

**Model Building**

We have list of following models:

•Logistic Regression : estimates probability of event occurring.

•Decision Tree : finds decision rules inferred from data in a tree manner.

•Random Forest : Combines results of multiple decision trees to single result.

•Naïve bayes : Classification model works based on bayes theorem.

•K-Nearest Neighbors : It uses the idea of similarity sometimes called distance, proximity, or closeness.

**Evolution Techniques**

•Accuracy score : ratio of correct predictions to the total prediction number.

•Precision : measures the proportion of positively predicted labels that are actually correct.

•Recall : represents the model's ability to correctly predict the positives out of actual positives.

•Confusion matrix : represents the prediction summary in matrix form.

**CONCLUSION**

In conclusion, we have analyzed important data patterns and trends from data using more advanced inbuilt methods and tools and applied different statistical methods. After data has been pre-processed the data is trained using machine learning classification techniques. The data is well trained with random forest and decision tree models with about 95% accuracy. There is potential use case left which can be implement next. That is the model can be integrated with different banks credit card processing systems so that decision can be made automatically.

**REPOSITORY/ARCHIVE**

**Github Link:** [**https://github.com/CharithaSai999/EmpiricalProject1**](https://github.com/CharithaSai999/EmpiricalProject1)

**RESOURCES**

1.Kaggle Credit Card Approval Prediction Competition

<https://www.kaggle.com/datasets/rikdifos/credit-card-approval-prediction/code>

2.Scikit-learn Machine Learning in Python

<https://scikit-learn.org/stable/>

3.Open AI GPT-3 Language Model

<https://platform.openai.com/docs/models/gpt-3>

**RELATED TUTORIALS**

**•**"Credit Card Fraud Detection Using Machine Learning" by M. Begum on Semantic Scholar

<https://www.semanticscholar.org/paper/Credit-Card-Fraud-Detection-using-Machine-Learning-Begum-Fatima/20e007b1a9ebdbcf3eb34e8b49bea45bde567ef3>

•"Predicting Credit Card Approvals" by Karolina Sowinska on Datacamp

<https://www.datacamp.com/projects/558>