# Title : Credit card data analysis with EDA and machine Learning approach

**Introduction:**

Delve into the realm of credit transactions with our EDA and Machine Learning approach. Uncover patterns, detect anomalies, and optimize decision-making. This project is your gateway to proactive financial strategies and enhanced risk management. Welcome to a concise yet impactful exploration of credit card data.

**Section A: Data analysis approach**

1. **What approach are you going to take in order to prove or disprove your hypothesis?**

**Data analysis approach**

By carrying out multivariate analysis Data analysis alongside visualization using seaborn, matplotlib, plotly and pandas plot function can help me find whether a given assumption or hypothesis is true or not.

**Hypothesis test**

By carrying out hypothesis tests I can check whether my initial hypothesis is True or not.

Here are a few hypotheses that I assumed beforehand with their outcome that I got through different analysis namely Hypothesis tests, data analysis and model comparison and evaluation.

**Hypothesis check-**

**Null hypothesis 1 - owning property and car would definitely have a positive effect on credit card application approved**

result - null hypothesis is rejected as p value is greater than alpha value which indicates that there is no statistically significant relation between property and label column & car owner and label column

**Null hypothesis 2 -gender would have no effect on credit approval decision**

result - null hypothesis is rejected as p value is less than alpha value it indicates that there is a significant relation between gender and target column.

**Null hypothesis 3-education column would definitely have an effect on credit approval decision**

result - null hypothesis is not rejected as p value is less than alpha which indicates there is a significant relation between education and target column

**Hypothesis check-Model comparison**

**Hypothesis**

Random forest classifier would perform much better than Decision tree classifier, KNN, SVM and Logistic regression

**Result**

Random forest algorithm is actually way better than other specified models in hypothesis as validated by model comparison.

1. **What feature engineering techniques will be relevant to your project?**

1-Feature Engineering - missing value imputation

- Identifying and treating outliers

-Handling categorical features(encoding)

-Feature scaling

2-Feature selection

3-feature construction

1. **Justifying the Data Analysis Approach :**

Before getting into the modelling stage, one has to have an in-depth understanding of data and patterns hidden within, that also includes a look through of data before getting into the analysis part.

I have gone through a series of steps in my Data analysis and even before getting into the pandas data analysis I have had a detailed go through on my data on excel where I have manually examined the data which is to be done to have an understanding before going into analysis.

**Aim of my Analysis**

* **To understand the data-** 
  + having a look into the data by checking its shape, presence of null values, duplicate rows etc.
* **Identifying patterns in data-**
  + Carrying out univariate and multivariate analysis along with visualizations to make patterns within data visible.
* **Testing whether a given assumption is True or not –**
  + There are many assumptions we take for granted but a deep dive into data can disprove such assumptions and can surface such details or anomalies**.**

**Setting an aim prior to analysis provides me with a flow to go through my analysis in a way that can help me the best way I can go from understanding to surfacing patterns in data through my analysis.**

1. **Identify important patterns in your data using the EDA approach to justify your findings.**

**Imbalanced data**

Data is highly imbalanced especially with respect to binary target variables in data where around 90% of observations belong to one of the categories, which is certainly going to affect the performance of the model.

**Education and Label**

Apart from higher education, Education doesn’t have an expected relation with the approval of credit card application, that is higher the education level doesn’t correspond to any direct relationship with application approval

**Employed-label**

One would expect employed people to have more likelihood of getting their application approved but as per data around 12% of unemployed have their application approved and for employed it drops to 10%.

This pattern might be the corollary of the fact that about 76% of unemployed people own property and only 63% are employed.

**(**although data have few unemployed as per numbers**)**

**Employed days -Label**

This again disproves a basic assumption that more employed days would have more chances of being their application approved.

Analysis reveals people who are getting their application received have around 700 less average employed days than those whose application is rejected.

**Income-label**

People having their application approved have higher average income. Which is mostly on expected lines

**Car owner/property owner-label**

Being a car or property owner doesn’t isn’t directly related to getting credit card application approved as per data around 10% of property owners have their application approved which is 11% in case of non-property owner.

**Gender-income**

Analysis reveals men have higher average income(222k) than women(172k) that too by a large margin.

**Children- income**

People who have less children (less than or equals to 2) have higher average income.

**Other-**

1-**employed day column**

This feature contained many impossible values where employed days were around 365243 which amounts to around 1000 years and all such values are there for people who are currently unemployed, So I replaced all such values 0 as they aren’t employed currently.

2-Labels seem to be randomly distributed in the data, which is having a negative effect on EDA.

3-Dataset is too small to capture any clear pattern in data, a larger dataset could have been helpful. (small size of data may limit the generality of analysis findings)

4-Many features which could have been really important for the analysis perspective aren’t present in data.

**Section B: Machine learning approach**

**Method that I use for machine learning based predictions for credit card approval**

**1-Feature engineering**

1A-Feature construction

There are few columns where I could create new meaningful features from existing features that could be effective in performance of my model.

1B-Feature transformation

* + 1. Missing value imputation

There are many missing values in data which need to be treated. I am going to use various techniques to handle missing values like random imputation for both categorical and numerical variables, KNN imputer for numerical features and simple imputer with different strategies based on feature type and requirement.

* + 1. Outlier treatment

Outliers in data can have a negative effect on performance. I might go for ‘capping strategy’ rather than trimming in case outliers are a considerable part of data, trimming which could result in loss of valuable data.

* + 1. Feature scaling

Feature scaling would ensure that all the features are on the same scale.

**2-Encoding Categorical columns**

Since machine learning algorithms need input data to be in numerical form encoding categorical columns becomes essential

**3-Model building**

1. Splitting the data into training and test set either using hold out or cross validation method
2. Training my data using different base models namely logistic regression, KNN, SVM, naïve bayes, Decision Tree, Random Forest, gradient boosting, voting classifier
3. After training these models I will use hyperparameter tuning to find the best set of parameters for each model for my desired classification metric.

**4-Model evaluation and selection**

After having all required metrics for different models and ensemble models (like gradient boosting, voting classifier, stacking etc) I will compare different models based on my desired metrics in tandem with other relevant metrics to get compare all models and select the best one out

**4B-Please justify the most appropriate model.**

Most appropriate model among all my models would be voting classifier with KNN classifier, decision tree , random forest, gradient boosting classifier and XG boost classifier as base models.

**Accuracy Metric**

As our model predicts whether one’s credit card application to be accepted or rejected, type 1 errors should be minimized, type 1 error would be harmful for business and would incur losses, along with that we also need to focus on type 2 errors and ensure that our model predicts the acceptance correctly that is we need to minimize type 2 error it will help in company expanding its customer base with many reliable clients.

Therefore in the context of the given problem I am focusing on recall as my accuracy metric and that too for both outcomes (0,1), apart from recall, f1 score is also a reliable metric here which takes both precision and recall in account.

**Justification for most appropriate model**

1. The model has a recall score of almost 1, that is it predicts all 0 outcomes (rejection of application) correctly and there are very low chances of any type1 error (false positive).
2. The model has a recall of around 50 percent for all 1 outcome (acceptance of application). 50% may seem less but with such small and imbalanced data getting a recall score of 1 for all 0 outcomes and around 50% for 1 outcome was the best combination that is achieved through this voting classifier model.

**4C-Please perform necessary steps required to improve the accuracy of your model.**

**4D-Please compare all models (at least 4 models)**

*I have performed all necessary steps to improve accuracy of model*

**1-mitigating the effect of biased classification using**-

1. over sampling
2. assigning class weights

**2-Feature selection**

**3-Hyperparameter tuning along with cross validation**

**4-Training data with 10 different models along with multiple ensemble models**

**5-Model evaluation and model selection**

Selecting the best model of all based on my desired metrics