

Project Title: Credit Card Approval Prediction (Classification Task)

About the Dataset:

Credit score cards are a common risk control method in the financial industry. It uses personal information and data submitted by credit card applicants to predict the probability of future defaults and credit card borrowings. The bank is able to decide whether to issue a credit card to the applicant. Credit scores can objectively quantify the magnitude of risk.

Your goal is to predict the credit card score status based on the applicant's information.

Credit Scores to predict:

- 'C': Likely denotes a "Credit Approved" status.
- 'X': Possibly stands for "Accepted" or "Approved." These statuses indicate that the application was successful, so they are grouped under the same category.
- '0', '1', '2', '3', '4', '5': These statuses could represent varying levels of rejection or issues with the application. For example:
 - '0': Application not approved but no specific issue.
 - '1': Denial due to minor reasons, like insufficient credit history.
 - '2': More severe issues, like high debt-to-income ratio.
 - '3' to '5': Potentially escalating levels of rejection or other reasons like bankruptcy, delinquency, or fraud.

You should Predict these classes and map the result to the following two status:

- Approved
- Not Approved

Here is the following mapping:

'C': 0, # Approved
'X': 0, # Approved
'0': 1, # Not Approved
'1': 1, # Not Approved
'2': 1, # Not Approved
'3': 1, # Not Approved
'4': 1, # Not Approved
'5': 1 # Not Approved

Notice that you have two datasets (“credit_record”/”application_record”) that will be merged into one. The common column between the two datasets is the ID so join the two datasets by joining the application of a person with the credit score dataset based on the ID.

Deliverables:

- A Python file with your complete implemented pipeline.
- Report explaining in details your work.

Rules:

- 1- Don't share code outside of the team (any detected plagiarism for any portion of the code will be considered as 0 for the whole project mark).
- 2- Don't use external data.

Project Minimum Requirements

1- Report: Students report their final work via a report that explains every step implemented.

e.g., We approached the problem with the assumption that features A, B,C affect the response variable the most ... etc.

2-Data Exploration and statistical analysis: Before applying any preprocessing technique, explore the data first, apply sanity check and see what preprocessing techniques you want to apply for data cleaning. Apply EDA on ("credit_record"/"application_record") before merging them.

3- Preprocessing: Before building your models, you need to make sure that the dataset is clean and ready-to-use.

Decide what you need to do about the data:

- Check for null values
- Apply feature engineering
- Drop unnecessary columns
- Check for class imbalance
- Check for garbage values
- Check for duplicates
- Merge datasets
- Label Encoding
- Feature scaling

Note that these steps are not ordered. You should search for problems in the data and solve these issues in logical order.

4-Feature Selection: Apply Genetic Algorithms technique to select the most important features for your model.(Self Study)

Use Genetic Algorithms for choosing the best features for decision trees model. The same features can be used for other models.

5- Splitting Data (Train / Validation/Testing):

Split your data into 70% train, 15% validation and 15% Testing.

When applying feature selection using genetic algorithms, the fitness function is the accuracy on the model's validation data.

6- Model Selection & training: You **MUST** train these classification

Models (KNN-Decision Trees –MLP). You can optionally add any other model you want to implement.

7- Hyper-parameter tuning:

- Apply grid-search or random search.

8- Model evaluation: When you get the best parameters for each classifier, you should consider the “accuracy” evaluation metric for evaluating all the hyperparameter-tuned classifier’s performance. Mention which classifier gave the best performance.

Project Deadline:

Friday 27/12/2024 at 11:59 PM

Deliver your work in the “Project Description/Project Submission “on moodle.