**Dataset description**

Write in your own words (1-2 sentences) what this dataset is about.

1. Does your description specify features and instances?
2. Do you describe the source and reliability of the data?

The dataset consist of two tables, application\_record and credit\_record, which are connected by the feature “ID”. application\_record has 18 features “ID(Primary Key), code\_gender, flag\_own\_car, flag\_own\_realty, cnt\_children, amt\_income\_total, name\_income\_type, name\_education\_type, name\_fam\_status, name\_housing\_type, days\_birth, days\_employed, flag\_mobil, flag\_work\_phone, flag\_phone, flag\_email, occupation\_type, cnt\_fam\_member” and credit\_record has 3 features “ID(ref application\_record(ID)), *months\_balance, status*”.

Source of data is from a real bank without sensitive personal information.

However, the dataset has been worked with by many so we will consider it reliable. (?)

While the actual data source remains undisclosed with no identifiable sources to validate the data, we can reasonably regard the dataset as reliable. This is so as the dataset is used extensively by others within the field. Additionally, given the sensitive nature of credit records, there is a need to safeguard individual identities, and such masking may be required to ensure data integrity and uphold data ethics.

**Project Title**

Please be concise, relevant and descriptive.

PG7 Prediction of Good & Bad Credit Records (with Random Forests) (?)

**Motivation**

Explain why this project is interesting and important.

1. Does your motivation clearly describe a problem?

Financial Sector, differentiate whether to approve a person’s credit card/loans etc…

Credit Card Ghosting: People apply but don’t use

Benefit for financial institutions, giving the approval only to those who really need it and capable of repaying. “Bad loans” have an effect on bank and financial stability. Financial institutions can reduce financial losses due to unpaid debts given the accurate predictions.

1. Does it justify the problem’s significance? What are the benefits of addressing this problem? Who benefits from solving it?

Customers:

Customers will have low credit scores, banks will reject them for future applications. Customers don’t know the algorithm for credit score used by banks, don’t know their chances of being accepted/rejected

Bankers:

Bankers can use more explainable models to improve transparency

Bankers stand to gain more credit-worthy customers. They also gain revenue as a result.

**Statement of the Problem/Task**

A statement of the problem, issue, or task that you’re interested in studying.

Try to formulate the 2-4 key questions that you will answer in the project.

Please clearly state your project topic and domain.

1. Does the proposal outline a problem statement, issue, or task that the team is interested in studying?

Problem statement: Identify major features that affect credit score records.

Key Questions:

At what point does inactivity count as ghosting for credit card companies?

How do their assets (car, property, etc.) determine their credit score records?

Which feature is the **most importan**t/has the highest correlation to a good credit score.

[Factors for credit card approval](https://www.cnbc.com/select/what-issuers-look-at-on-credit-card-applications/)

1. Does it formulate a few (2 to 4) questions that the team proposes to address?

**General Approach**

A high-level description of the general approach you’ll use to address the questions. Sketch out what evidence you are planning to gather (e.g. how you can answer the questions through experiments on data). Survey on the current progress on the problem/task.

1. Does the proposal contain a high-level draft description of the general approach proposed to address the questions?

Data cleaning and preprocessing:

* Remove empty values in dataset (Indicated by question mark)
* Median/Mode imputation for missing values
* Encoding categorical data (such as code\_gender and flag\_own\_car columns, flag\_own\_realty)
* normalisation/ standardisation for data with large values

Look at the distribution of the dataset (EDA)

1. Does it include preliminary plans for evaluation, data gathering? I.e., how the team plans to answer the questions through experiments on data.

Split the data into training and testing.

Training Random forest/decision tree and predicting on test data.

Metric to judge using: accuracy, f1 score (for imbalanced data), ROC (if higher priority is given to approval than rejection)

Analyse and improve model

**Evaluation**

Include how you will evaluate your project. Propose what your team thinks is a satisfactory project outcome (C grade) and an excellent project outcome (A grade). Remember that performance is secondary to analysis and understanding.

1. Does the proposal contain a high-level draft description of the general approach proposed to address the questions?

Rate performance with accuracy on test data. Excellent project outcome (High accuracy score) and satisfactory project outcome (Lower accuracy score)

Secondly, we will also rate performance on f1 score due to the unbalanced nature of the dataset

We can also use a precision-recall curve to give better priority to TPR or FPR.

1. Does it include preliminary plans for evaluation, data gathering? I.e., how the team plans to answer the questions through experiments on data.

**Resources**

A list of resources you have/need to conduct the project. This includes additional reading, software, datasets, code(github link), etc., beyond your chosen dataset. Are these resources public? How are you planning to get these resources?

Kaggle Discussion Page for Dataset: <https://www.kaggle.com/datasets/rikdifos/credit-card-approval-prediction/discussion?select=application_record.csv>

Github

Libraries used: sklearn, numpy, pandas

1. Does the proposal give a short list of resources the team plans to use to execute the project (inclusive of readings, software, datasets, etc)?
2. Does the team describe any strategy for getting the resources?

**Schedule / Role Assignment**

A schedule of work indicating the dates by which you plan to complete components of the project. Make sure the schedule is plausible.

You may find that a table format with the remaining weeks of the course helpful to describe this goal.

1. A schedule indicating dates by which the team plans to complete the project components?

| Week | No. of people (I specify names ltr for the form) | Things to Finish |
| --- | --- | --- |
| Recess Week | Everyone | Preliminary research (Github and Kaggle notebooks) |
| 3 (i random later) | Cleaning up the data & Exploratory Data Analysis (EDA) |
| 3 | Feature Engineering (e.g. featuretools) |
| 3 | Training and testing the model |
| 3 | Evaluating performance (accuracy, F1 score, precision-recall curve) |
| Week 10-12 | everyone? | Finding ways to improve performance (ensemble training, etc.) |
| Week 13 | everyone? | Additional feedback and improvements from mentors, TAs and Profs |
| Reading Week and Exam Week 1 | Everyone ( each work on one section) | Work on presentation slides + rehearse + film presentation video |
| Exam Week 1 (3 Dec, Sunday) |  | Final Presentation submission |

1. An assignment of the team members to the deliverables?
2. Is the schedule feasible given the timeline, expertise and load of the team members?

Yes, it is feasible, as we have taken into account the team members’ schedules.

1. For projects with possibly too large a data source (e.g., Kaggle projects), does the team propose a way to scope the data or problem accordingly to make it feasible?

The dataset is manageable as it contains less than 200k rows? And the all the features are useful for analysis