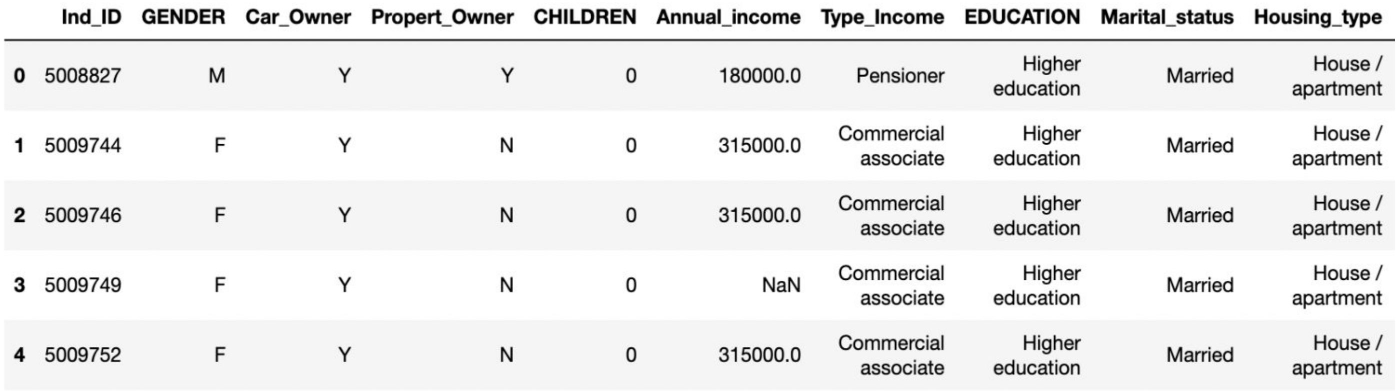
# Predictive Analysis of Credit Card Application Outcomes

## Project Deliverable 2

Sindhuja Rai, Joey Xia,
Mansi Narendra Singh, Kevin Taylor

# Introduction

- Financial institutions may use several factors to determine whether an applicant is accepted or rejected for a credit card.
- Our project is based on the Credit Card Approval Dataset and the key factors that contribute to the approval or denial of a credit card application.
- The dataset contains a number of numerical and categorical variables as seen below, as well as an application status encoded by 0 - “Approved” or 1 - “Rejected”.

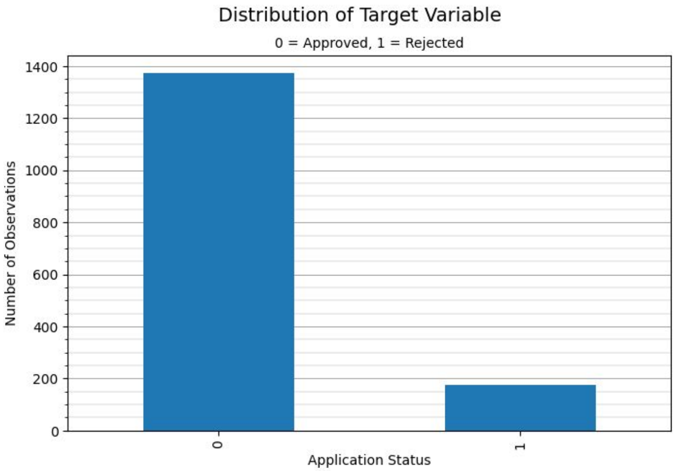


| | Ind_ID | GENDER | Car_Owner | Propert_Owner | CHILDREN | Annual_income | Type_Income | EDUCATION | Marital_status | Housing_type |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 5008827 | M | Y | Y | 0 | 180000.0 | Pensioner | Higher education | Married | House / apartment |
| **1** | 5009744 | F | Y | N | 0 | 315000.0 | Commercial associate | Higher education | Married | House / apartment |
| **2** | 5009746 | F | Y | N | 0 | 315000.0 | Commercial associate | Higher education | Married | House / apartment |
| **3** | 5009749 | F | Y | N | 0 | NaN | Commercial associate | Higher education | Married | House / apartment |
| **4** | 5009752 | F | Y | N | 0 | 315000.0 | Commercial associate | Higher education | Married | House / apartment |

# Initial Data Exploration

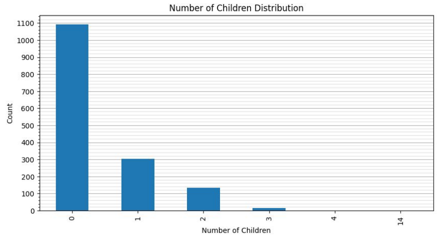
### *Distribution of the Target Variable*

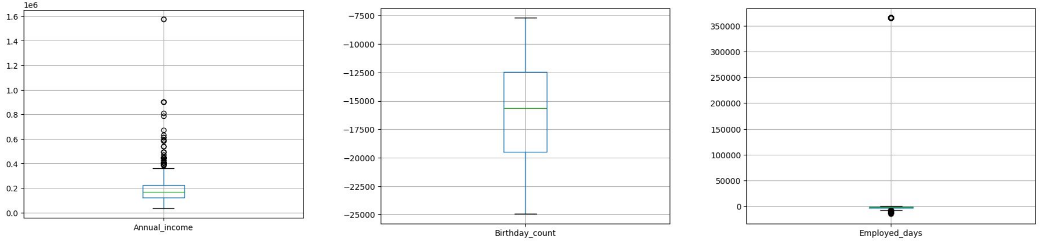
The dataset is imbalanced, containing 175 (88.7%) records of 1 - “Rejected” and 1,373 (11.3%) records marked 0 - “Accepted”



### *Distribution of Numerical Variables (Examples)*

*The applicant with 14 children can be considered as a fault in the data or an outlier since it is highly unlikely.*





*Other numerical variables ie) annual income, birthday count (age in days), or employed days will likely need to be binned or log-transformed.*

# Initial Data Exploration

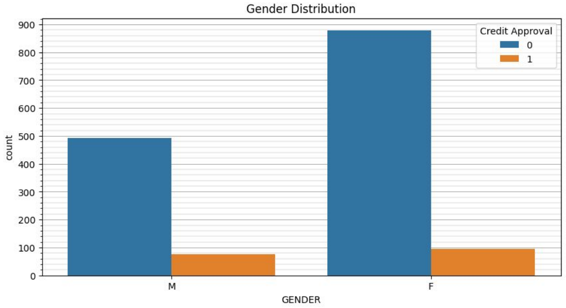
***Distribution of Categorical Variables w.r.t. the Target Variable (Examples)***

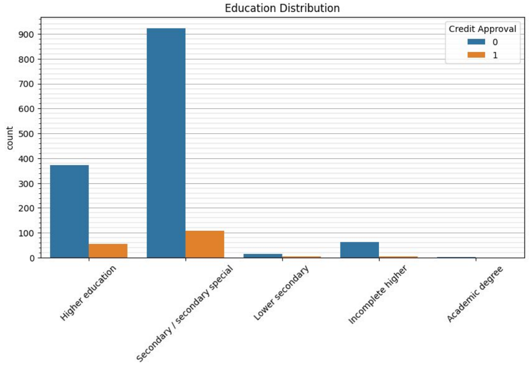
**Gender** - Most applicants are female.

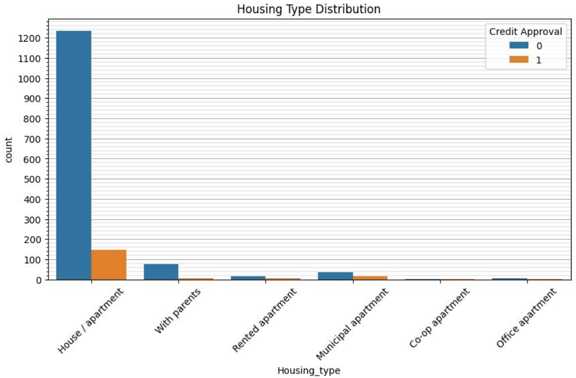
**Housing Type** - Most applicants live in apartments.

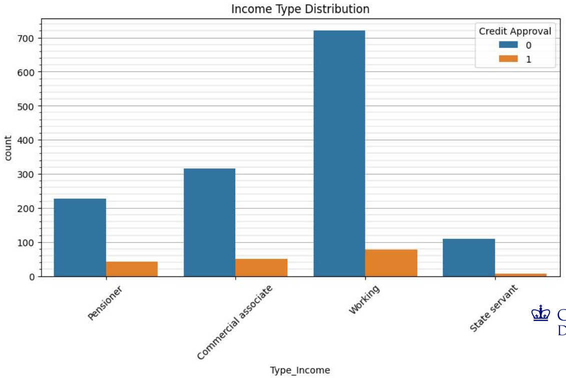
**Education Level** - Most applicants' highest level of education is secondary education.

**Income Type** - Most applicants are working class.


Gender Distribution

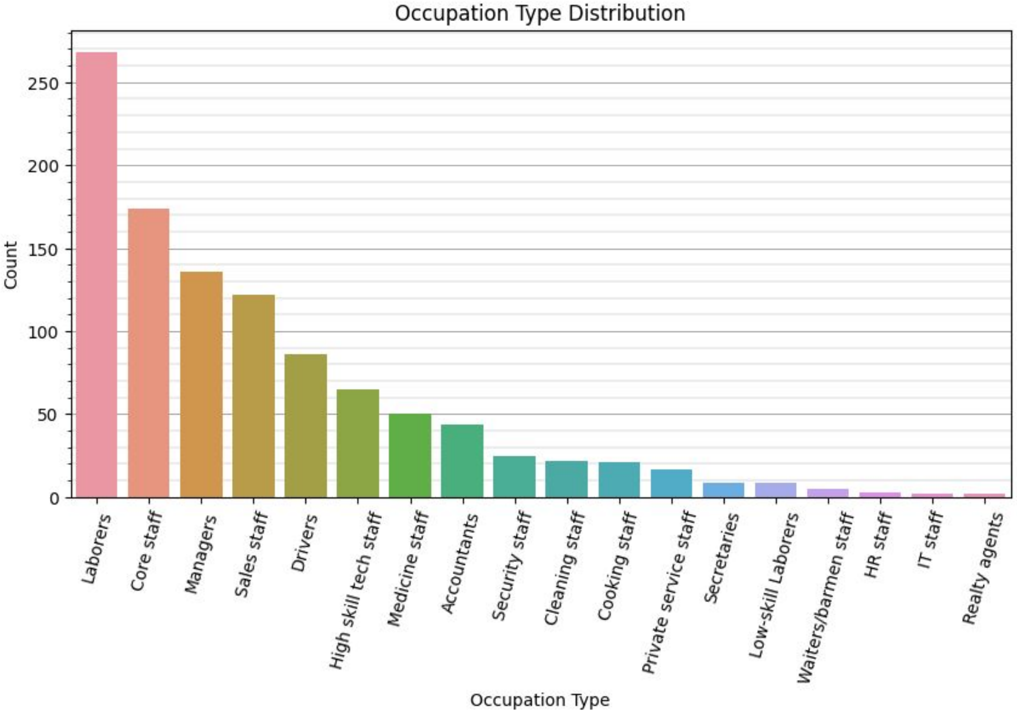

Education Distribution


Housing Type Distribution


Income Type Distribution

# Initial Data Exploration

***Distribution of Categorical Variables***



*Some variables, ie) Occupation Type, have many categories and will need to be represented with target encoding.*

# Data Cleaning

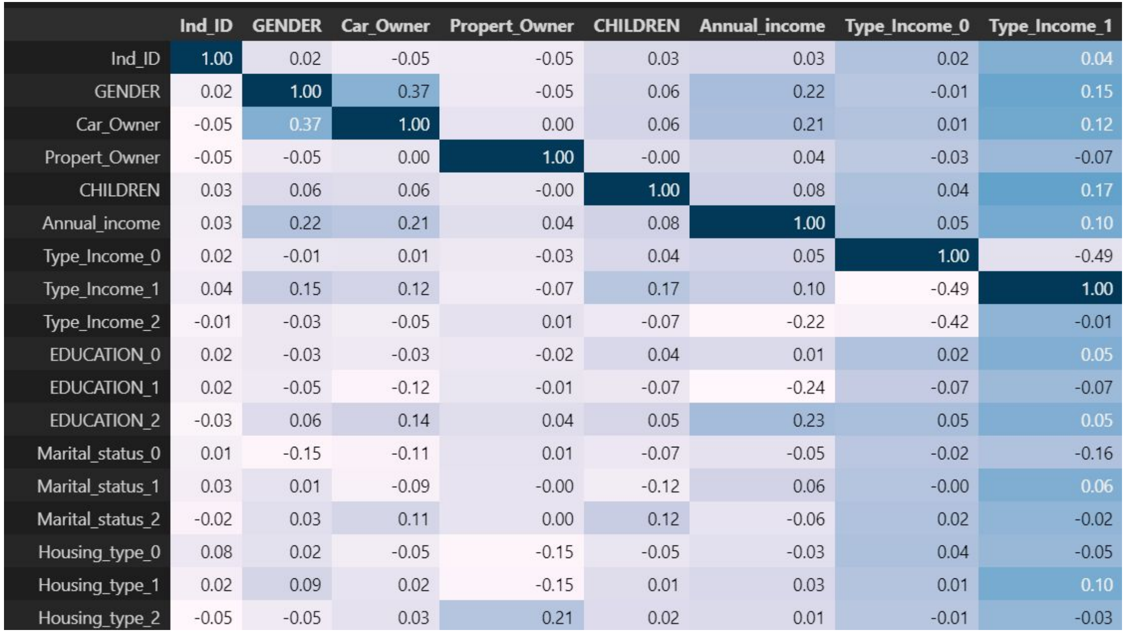
<u>Missing entries were handled in the following ways:</u>

- There are 4 columns with missing entries.
- Assigned 'Unknown' to 7 missing entries in the column Gender and 488 missing values in Type of Occupation to maintain data integrity - if these features are unknown it may affect the application status.
- Missing values in Annual Income and Birthday Count was imputed with the median to avoid skewing from the outliers.

<u>Encoded Categorical Variables in the following ways:</u>

- Binary variables such as Gender, Car Owner, Property Owner, Employment Status were encoded using ordinal encoding.
- For variables with 4-5 categories such as Type of Income, Marital Status, Housing Type and Education, binary encoding is used because it doesn’t impose an ordinal relationship and it reduces the dimensionality compared to one-hot encoding.
- For Type of Occupation, target encoding is used since it has high cardinality and can capture more information about the target variable within the feature which could be helpful for the predictive model.

# Data Cleaning



| | Ind_ID | GENDER | Car_Owner | Propert_Owner | CHILDREN | Annual_income | Type_Income_0 | Type_Income_1 |
|---|---|---|---|---|---|---|---|---|
| Ind_ID | 1.00 | 0.02 | -0.05 | -0.05 | 0.03 | 0.03 | 0.02 | 0.04 |
| GENDER | 0.02 | 1.00 | 0.37 | -0.05 | 0.06 | 0.22 | -0.01 | 0.15 |
| Car_Owner | -0.05 | 0.37 | 1.00 | 0.00 | 0.06 | 0.21 | 0.01 | 0.12 |
| Propert_Owner | -0.05 | -0.05 | 0.00 | 1.00 | -0.00 | 0.04 | -0.03 | -0.07 |
| CHILDREN | 0.03 | 0.06 | 0.06 | -0.00 | 1.00 | 0.08 | 0.04 | 0.17 |
| Annual_income | 0.03 | 0.22 | 0.21 | 0.04 | 0.08 | 1.00 | 0.05 | 0.10 |
| Type_Income_0 | 0.02 | -0.01 | 0.01 | -0.03 | 0.04 | 0.05 | 1.00 | -0.49 |
| Type_Income_1 | 0.04 | 0.15 | 0.12 | -0.07 | 0.17 | 0.10 | -0.49 | 1.00 |
| Type_Income_2 | -0.01 | -0.03 | -0.05 | 0.01 | -0.07 | -0.22 | -0.42 | -0.01 |
| EDUCATION_0 | 0.02 | -0.03 | -0.03 | -0.02 | 0.04 | 0.01 | 0.02 | 0.05 |
| EDUCATION_1 | 0.02 | -0.05 | -0.12 | -0.01 | -0.07 | -0.24 | -0.07 | -0.07 |
| EDUCATION_2 | -0.03 | 0.06 | 0.14 | 0.04 | 0.05 | 0.23 | 0.05 | 0.05 |
| Marital_status_0 | 0.01 | -0.15 | -0.11 | 0.01 | -0.07 | -0.05 | -0.02 | -0.16 |
| Marital_status_1 | 0.03 | 0.01 | -0.09 | -0.00 | -0.12 | 0.06 | -0.00 | 0.06 |
| Marital_status_2 | -0.02 | 0.03 | 0.11 | 0.00 | 0.12 | -0.06 | 0.02 | -0.02 |
| Housing_type_0 | 0.08 | 0.02 | -0.05 | -0.15 | -0.05 | -0.03 | 0.04 | -0.05 |
| Housing_type_1 | 0.02 | 0.09 | 0.02 | -0.15 | 0.01 | 0.03 | 0.01 | 0.10 |
| Housing_type_2 | -0.05 | -0.05 | 0.03 | 0.21 | 0.02 | 0.01 | -0.01 | -0.03 |

...

...

Correlation among each of the variables was calculated.

Variables *CHILDREN* and *Family_Members* understandably have a correlation of 0.89 - one of the variables should be dropped.

The unique application ID was also dropped from each row.

# Data Sampling

- The dataset exhibits an imbalance with 88.7% of credit card applications approved and 11.3% rejected.
- To maintain proportional representation of  approved and rejected applications in all data subsets, stratified sampling should be used.
- The dataset was divided using stratified sampling with 80% as the development set (which will be used for training / validation) and 20% as test set for model evaluation.
- Used random_state parameter during the split to ensure reproducibility of the results.
- During model development and hyperparameter tuning, due to the imbalanced nature of the data, techniques such as Oversampling/Undersampling, Ensemble Resampling, SMOTE or class weights may be utilized.

# Proposed Models

The goal is to perform binary classification on the dataset. Each models to be considered, whether the model requires variable encoding, and some of the hyperparameters in each, are detailed below:

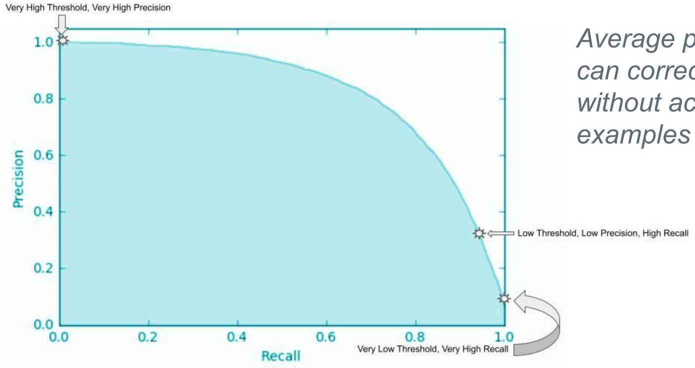
| Machine Learning Model | Encoding Required | Hyperparameters |
|---|---|---|
| Logistic Regression | Yes | L1, L2 penalties as applicable |
| Support Vector Machines | Yes | Primal/Dual, Relaxation parameters, etc. |
| Decision Tree | No | Min samples split/leaf, max depth, max leaves, impurity measure, etc. |
| Random Forest | No | Number estimators, decision tree parameters, etc. |
| Gradient-Boosted Trees | No | Number estimators, base estimator, etc. |
| Neural Network | Yes | Network architecture, learning rate, activation function, etc. |

# Proposed Metrics

- The dataset is imbalanced - the minority class is class 1 - “Rejected.”
- Because of the imbalanced dataset, the models should be evaluated on the minority class, or macro averages between the majority and minority class. Precision and recall should be used rather than accuracy.
- The nature of the data also indicates a false positive (Classified “Rejected” when actually “Accepted”) has a high consequence - for this reason, metrics sensitive to false positives should be chosen.

**Proposed Metrics for Model Evaluation:**

- Minority Class Average Precision
- Macro-Weighted Precision



*Average precision indicates whether the model can correctly identify all the positive examples without accidentally marking too many negative examples as positive.*