Credit Card Approval Prediction

Presented by: DataSweeper Technolgy Inc

Jane Huang
Lucas Chandra
Binoy Luckoo
Samir Rifi

September, 2021

DataSweeper Technology Inc (DTI)

DTI added Providence Bank, located in the Bahamas, to its client portfolio.

The bank wants to minimize the risks involved in its credit card client portfolio.

DTI's first mandate is to develop a machine learning model that can predict whether a credit card applicant will be approved or denied and identify the applicant attributes that have a major impact on the decision.

The decision of approving a credit card is mainly dependent on the personal and financial background of the applicant. Factors like, age, gender, income, employment status, credit history and other attributes all carry weight in the approval decision.

Question (it refers to the team wants to answer with the data)

1. Based on the dataset, what are the standard requirements for an individual to be approved for a credit card?

2. Can the model minimize the following risks:

- Loss from not approving the good applicant
- Loss resulting from approving a non-credit worthy candidate

Project Plan

DTI assigned a team of four Data Scientists to this project with Lucas C. as the lead.

The project plan is as follows:

1. Pre-Analysis of the data to decide which technologies to use
2. Pre-processing of two datasets provided by Providence Bank
3. Analysis of the demographics of the datasets
4. Run different Machine Learning models on the dataset
5. Decide which Machine Learning model is best suited for the bank
6. Present findings and recommendations to the bank

Dataset

The dataset used for the analysis is from kaggle and can be accessed at **Credit Card Approval Prediction**

The Dataset contains two files:

1. Demographics & application data - "application_record.csv"

   This data has been provided by the applicants at the time of the credit card application. It contains demographic information including gender, car & real estate ownership, income level, education, occupation, marital status, contact information.

2. Credit Bureau data - "credit_record.csv"

   Data obtained from the credit bureau showing payment experience and the date of the last data extraction.

Technology Stack

Project: Credit Card Approval Prediction

Technology Stack

Dataset: downloaded from kaggle https://www.kaggle.com/rikdifos/credit-card-approval-prediction/code

Exploratory Data Analysis: jupyter, python, matplotlib, seaborn

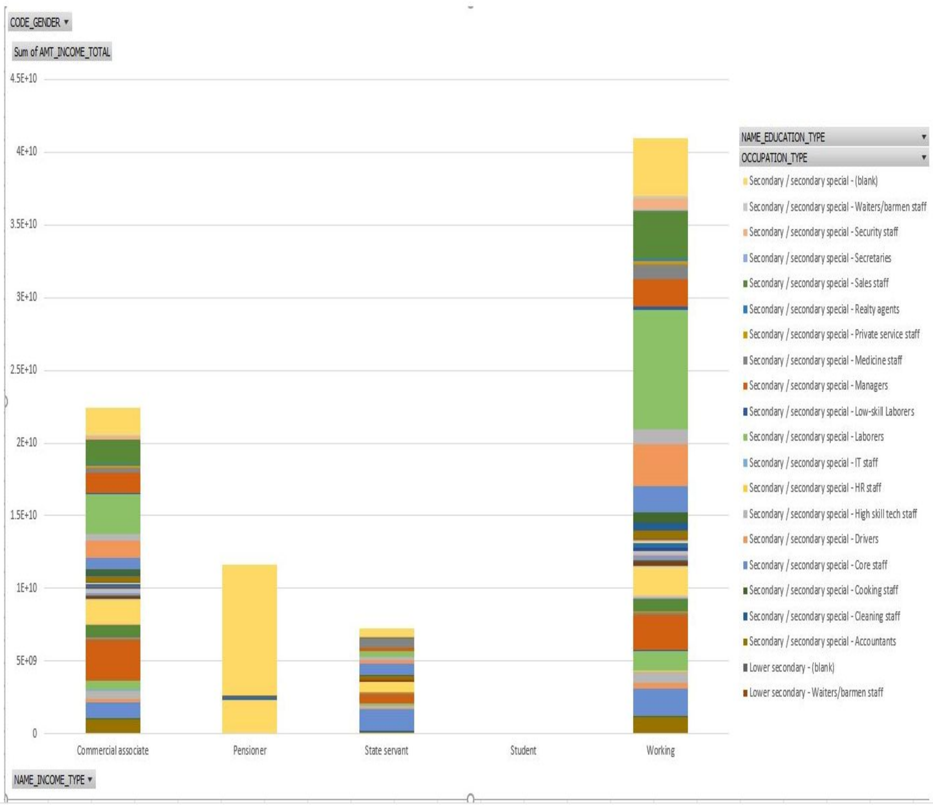
Database, Data Wrangling & Feature Engineering: pandas, PostgreSQL, NumPy

Machine Learning Pipeline: Scikit learn

Dashboard Presentation: Google Slides, Html, JavaScript, CSS

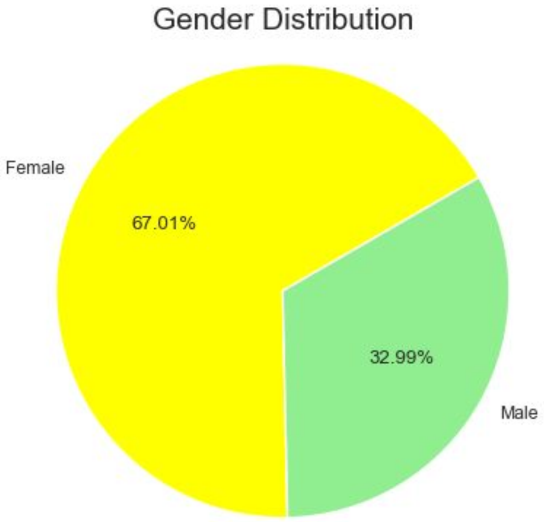
DataSweeper_Project

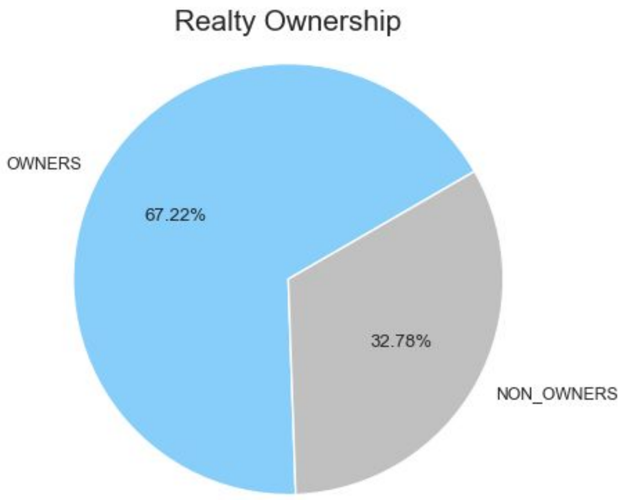
Data Exploration



DATASET DEMOGRAPHICS
Gender Distribution & Realty Ownership

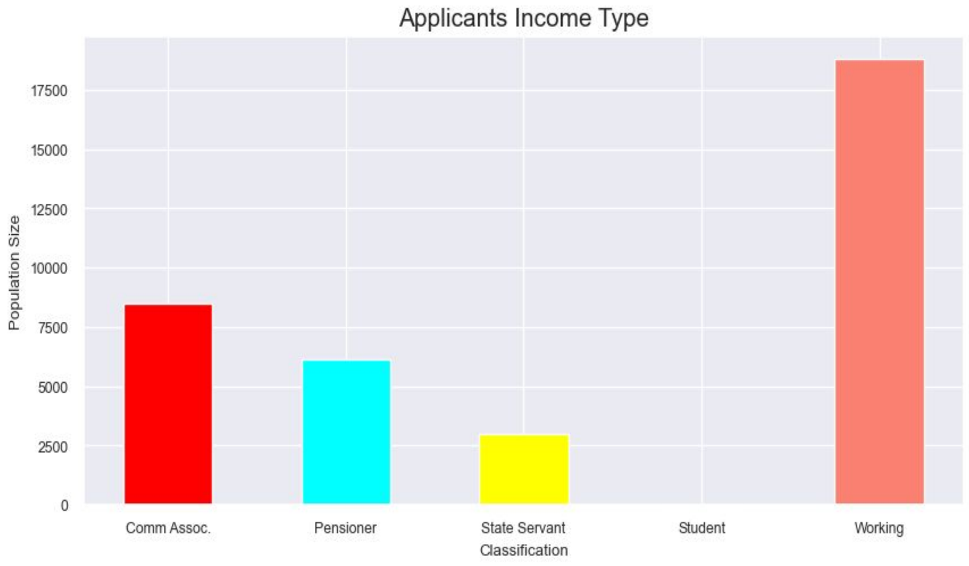
These charts show the gender distribution and realty ownership status of all applicants in the datasets being used for the models.

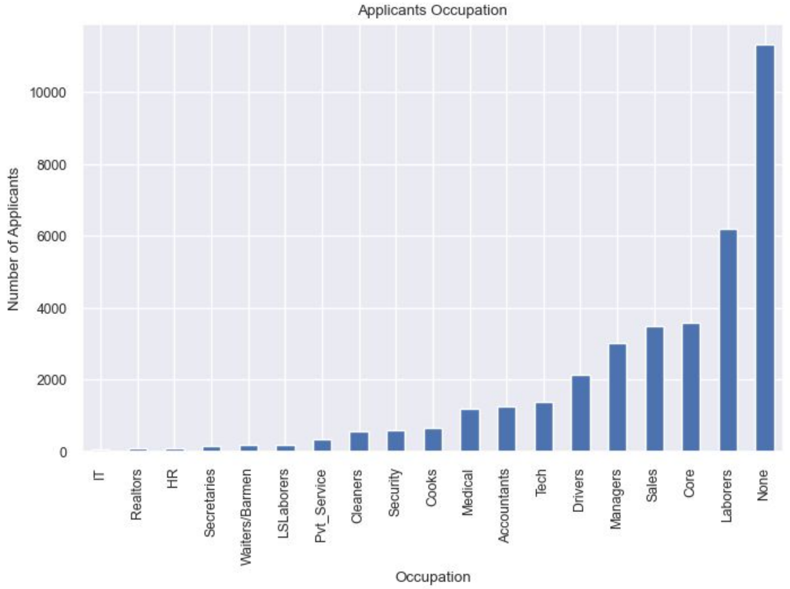




DATASET DEMOGRAPHICS
Applicants Income Type & Occupation

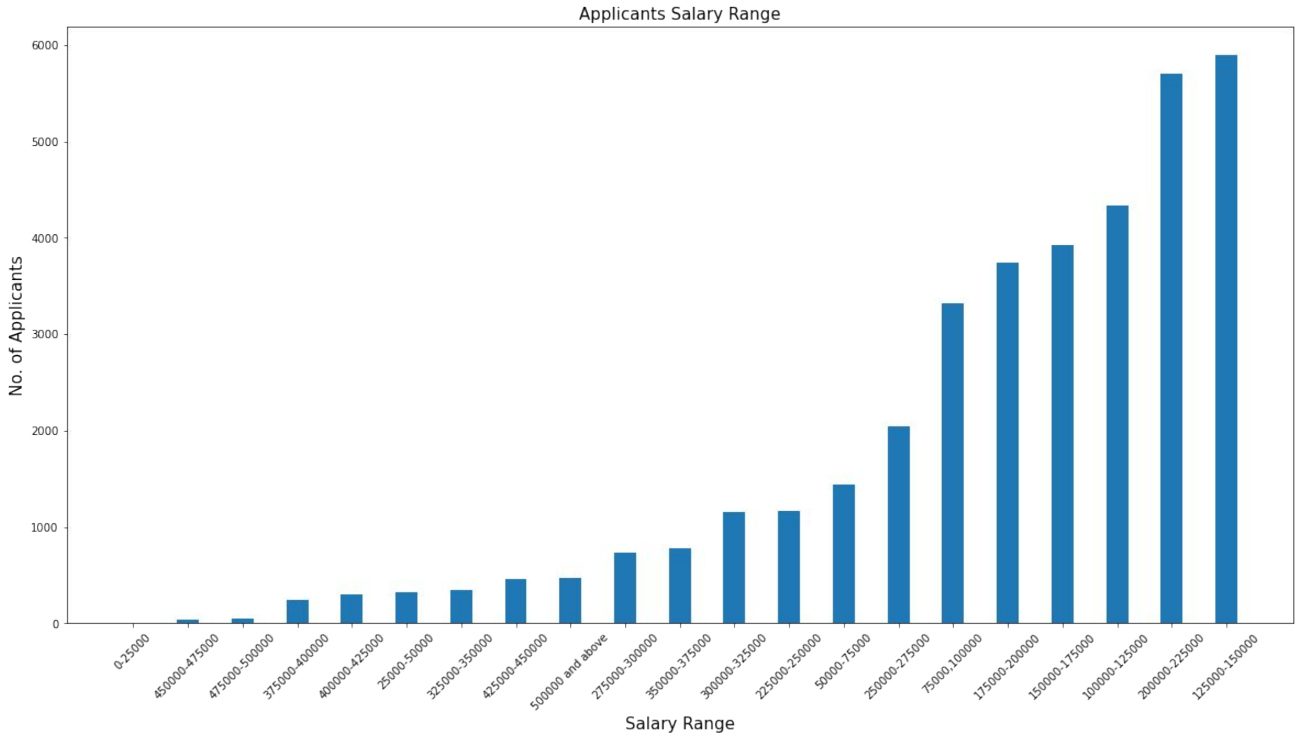
Applicants income type and occupation are displayed in the following charts





DATASET DEMOGRAPHICS
Applicants Salary Range

The datasets provided have a high number of applicants skewed towards high salaries.



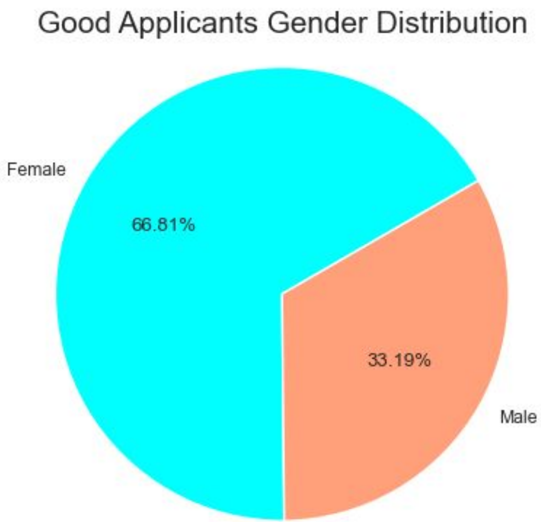
MACHINE LEARNING

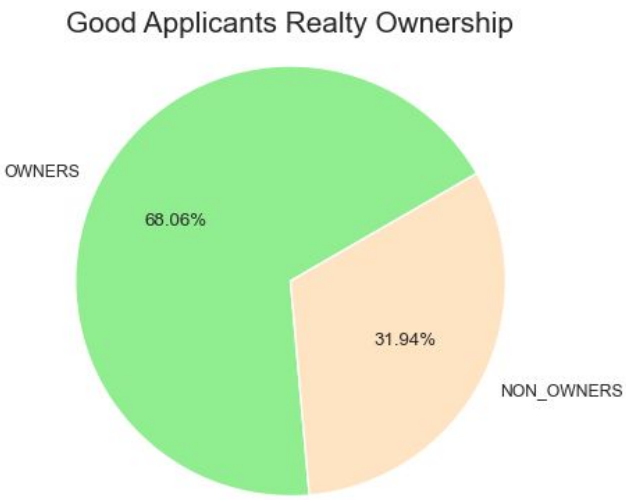
Data Processing	Clean the data Joins → pgAdmin Merge → Pandas
Features	Random Oversampling SVM Decision Tree Random Forest
Training & Testing Sets	Y value → X value →
Model Choice	?
Accuracy Scores	Training → Testing →

Data Analysis

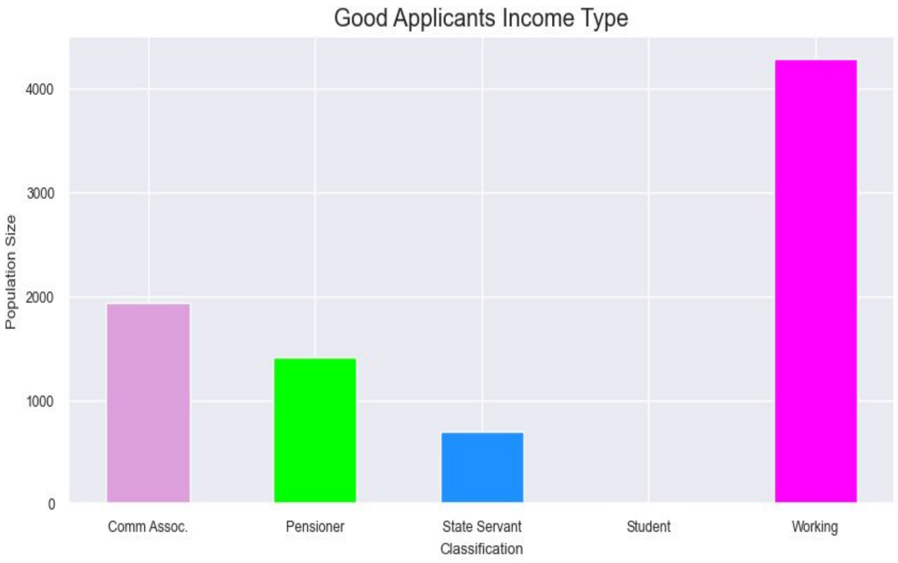
GOOD APPLICANTS DEMOGRAPHICS
Gender Distribution & Realty Ownership

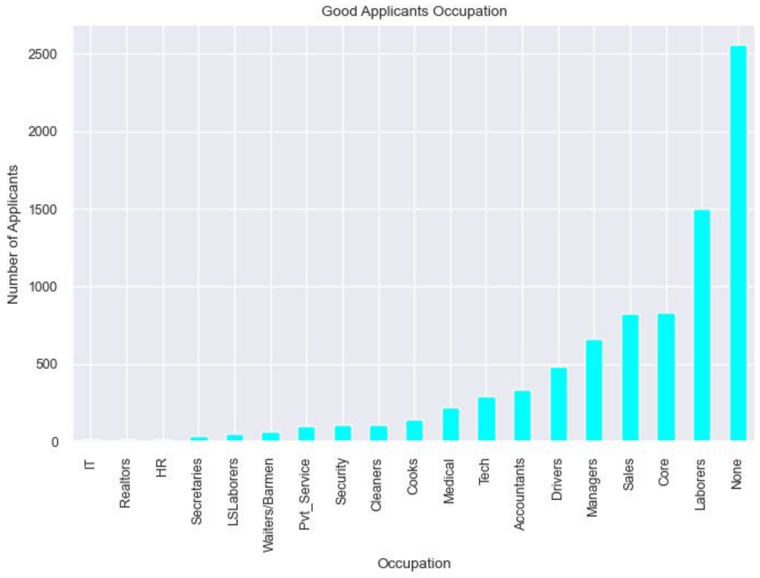
An analysis of the "good applicants" show that the distribution follows the same demographics as the whole population. This is depicted in the following charts.





GOOD APPLICANTS DEMOGRAPHICS
Income Type & Occupation





DATABASE

Data Processing	Extract & Transform: Jupyter Notebook, Python, Pandas
Database Setup	AWS RDS, PostgreSQL
Table Joins	Database: pgAdmin

Machine Learning Models

Data Processing

Extract & Transform:

Jupyter Notebook, Python, Pandas

Machine Learning Models

The DTI team cleaned the data and processed it in different Machine Learning models to determine which model best fits the requirements of the bank.

Each model is evaluated based on:

- Confusion Matrix - performance measurement showing 4 quadrants
    1. True Negative: prediction indicates “Bad” applicant and applicant is actually “Bad”
    2. False Positive (referred to as a Type 1 Error): prediction indicates “Good” applicant and applicant is actually “Bad”
    3. False Negative (referred to as a Type 2 Error): prediction indicates “Bad” and actual applicant is actually “Good”
    4. True Positive: prediction indicates “Good” applicant and applicant is actually “Good”

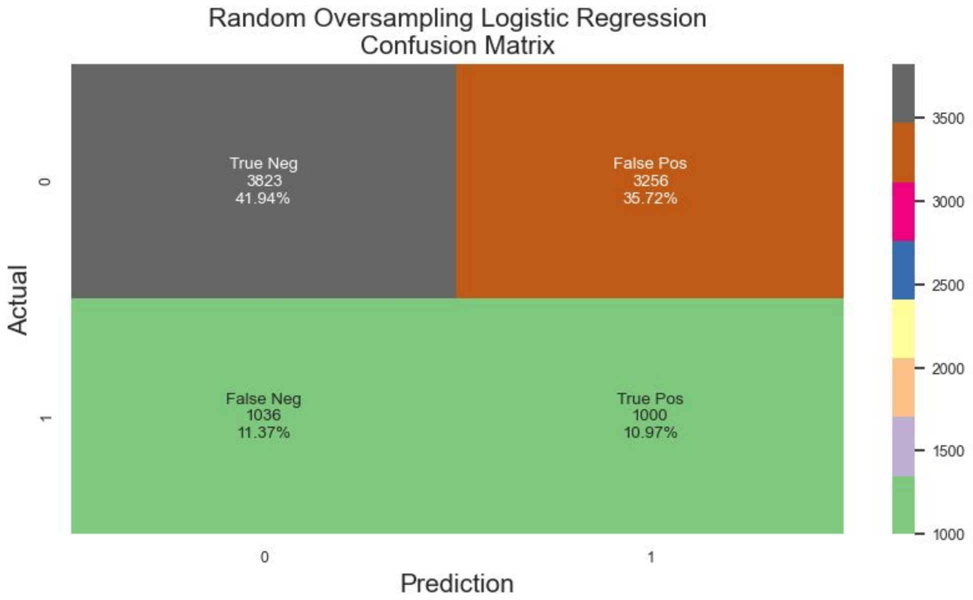
- Classification Reports -
    - Precision - for all the applicants classified as “Good” or “Bad” how many are actually “Good” or “Bad” respectively
    - Recall - from the “Good”, what percentage were predicted correctly
    - Accuracy - from the applicants classifications, what percentage were predicted correctly
    - F1-Score - a combination of precision and recall. A high F1 score is an indication that the predictions have low quantities of false “Good” and false “Bad”

The following charts is an illustration of the above metrics for each model.

RANDOM OVERSAMPLING
Logistic Regression

Classification Report

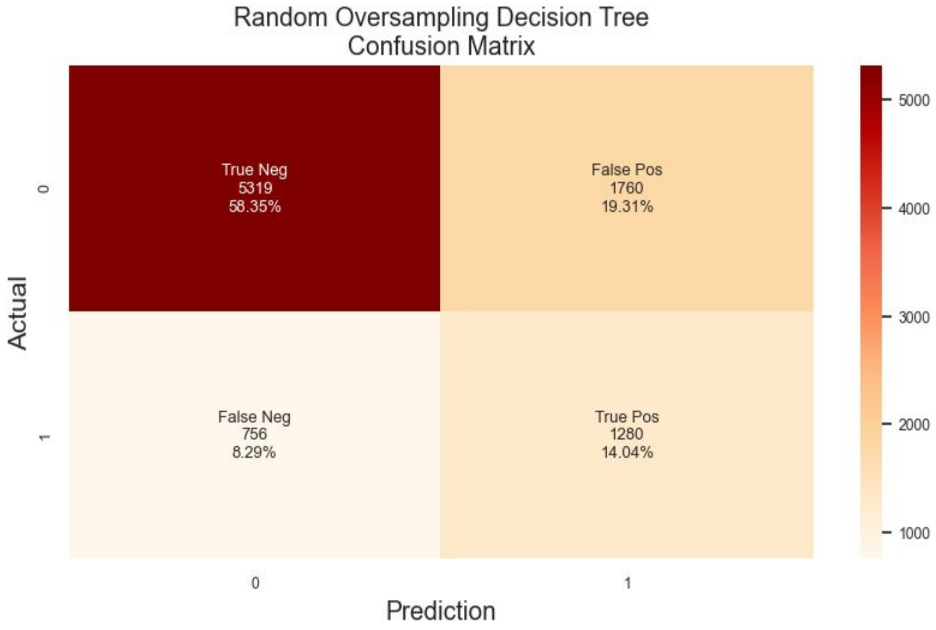
	Precision	Recall	f1-Score	Support
0	0.7868	0.5400	0.6405	7079
1	0.2350	0.4912	0.3179	2036
accuracy	0.5291	0.5291	0.5291	0
macro avg	0.5109	0.5156	0.4792	9115
weighted avg	0.6635	0.5291	0.5684	9115



RANDOM OVERSAMPLING
Decision Tree

Classification Report

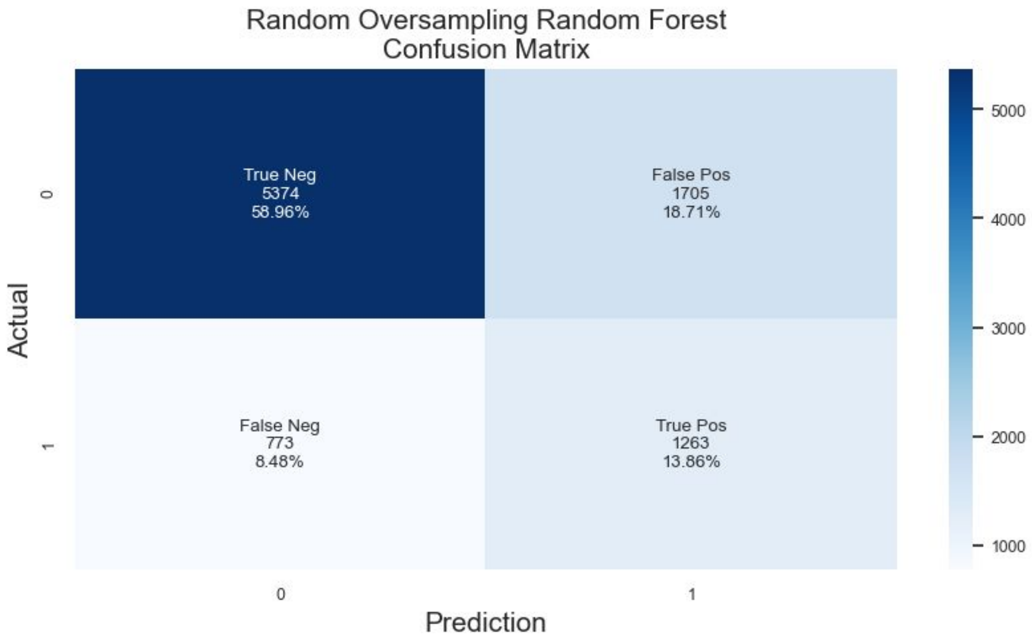
	Precision	Recall	f1-Score	Support
0	0.8756	0.7514	0.8087	7079
1	0.4211	0.6287	0.5043	2036
accuracy	0.7240	0.7240	0.7240	0
macro avg	0.6483	0.6900	0.6565	9115
weighted avg	0.7740	0.7240	0.7407	9115



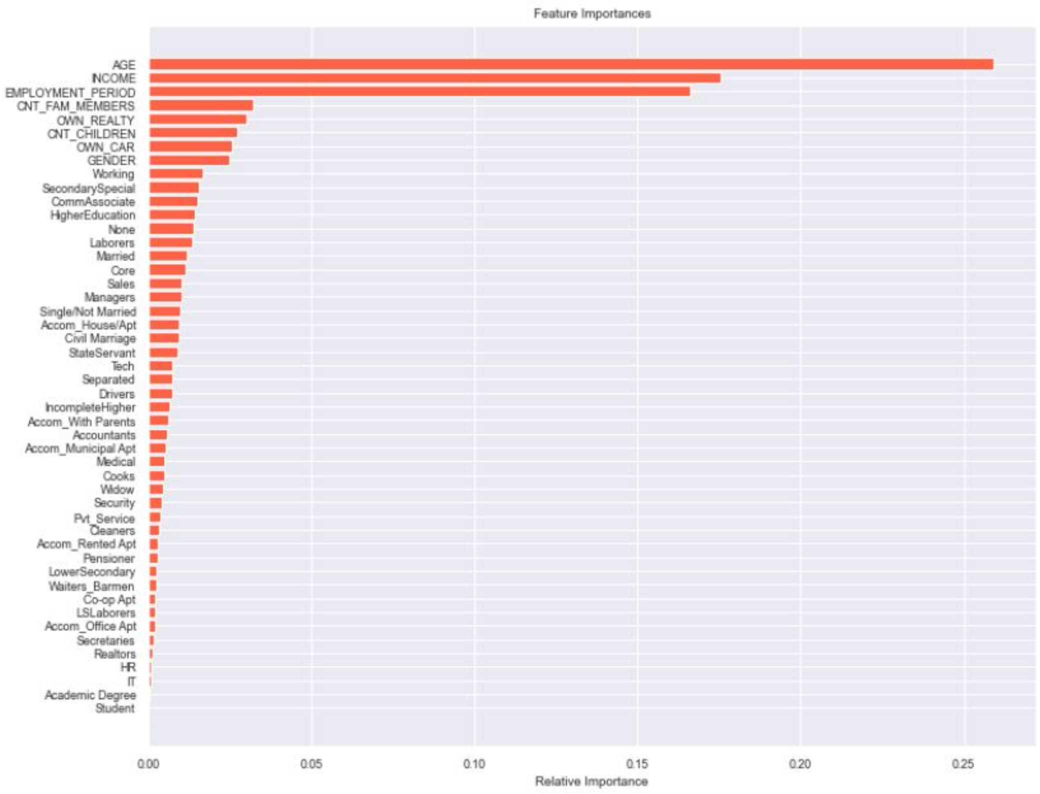
RANDOM OVERSAMPLING
Random Forest

Classification Report

	Precision	Recall	f1-Score	Support
0	0.8742	0.7591	0.8126	7079
1	0.4255	0.6203	0.5048	2036
accuracy	0.7281	0.7281	0.7281	0
macro avg	0.6499	0.6897	0.6587	9115
weighted avg	0.7740	0.7281	0.7439	9115



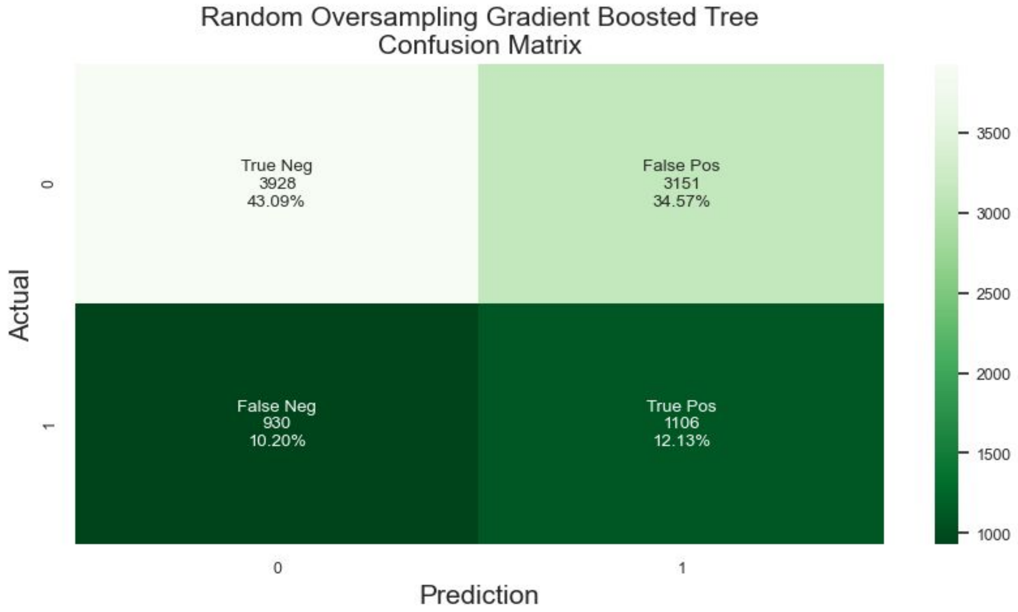
RANDOM OVERSAMPLING
Random Forest - Feature Importance



RANDOM OVERSAMPLING
Gradient Boosted Tree

Classification Report

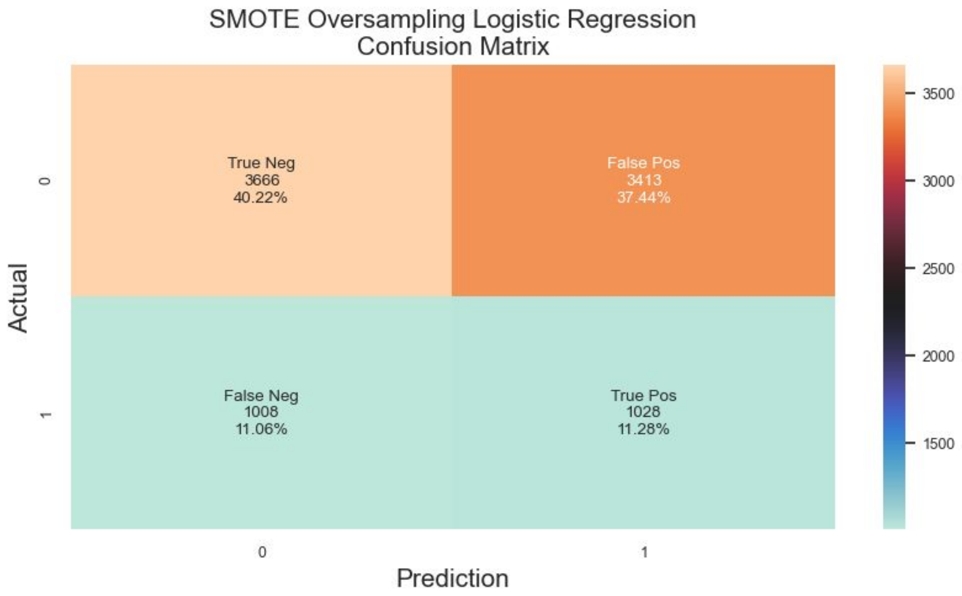
	Precision	Recall	f1-Score	Support
0	0.8742	0.7591	0.8126	7079
1	0.4255	0.6203	0.5048	2036
accuracy	0.7281	0.7281	0.7281	0
macro avg	0.6499	0.6897	0.6587	9115
weighted avg	0.7740	0.7281	0.7439	9115



SMOTE OVERSAMPLING
Logistic Regression

Classification Report

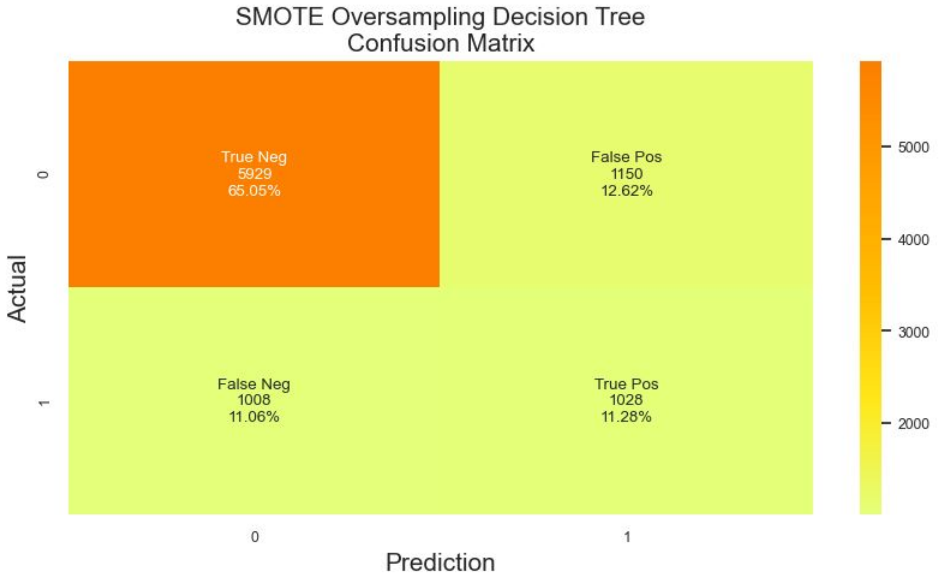
	Precision	Recall	f1-Score	Support
0	0.7843	0.5179	0.6238	7079
1	0.2315	0.5049	0.3174	2036
accuracy	0.5150	0.5150	0.5150	0
macro avg	0.5079	0.5114	0.4706	9115
weighted avg	0.6608	0.5150	0.5554	9115



SMOTE OVERSAMPLING
Decision Tree

Classification Report

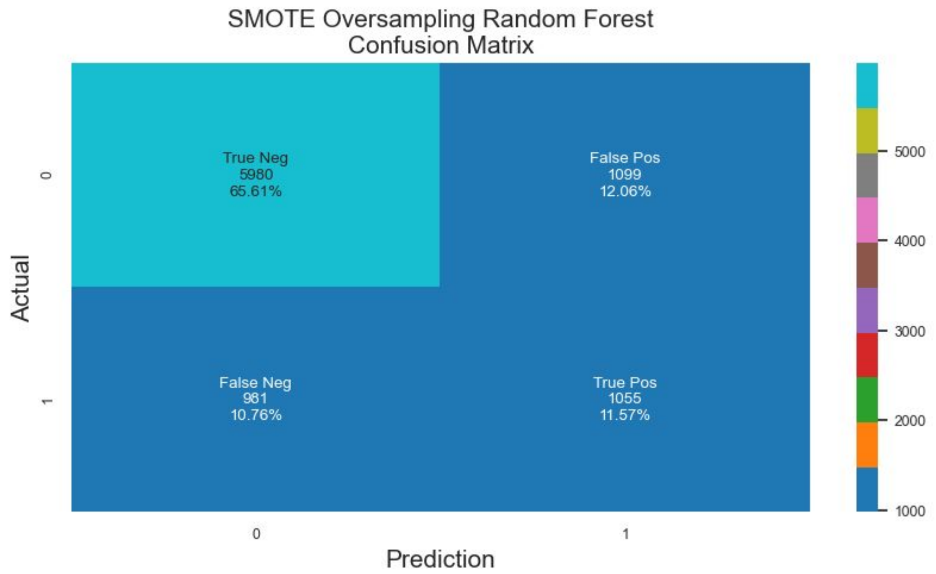
	Precision	Recall	f1-Score	Support
0	0.8547	0.8375	0.8460	7079
1	0.4720	0.5049	0.4879	2036
accuracy	0.7632	0.7632	0.7632	0
macro avg	0.6633	0.6712	0.6670	9115
weighted avg	0.7692	0.7632	0.7660	9115



SMOTE OVERSAMPLING
Random Forest

Classification Report

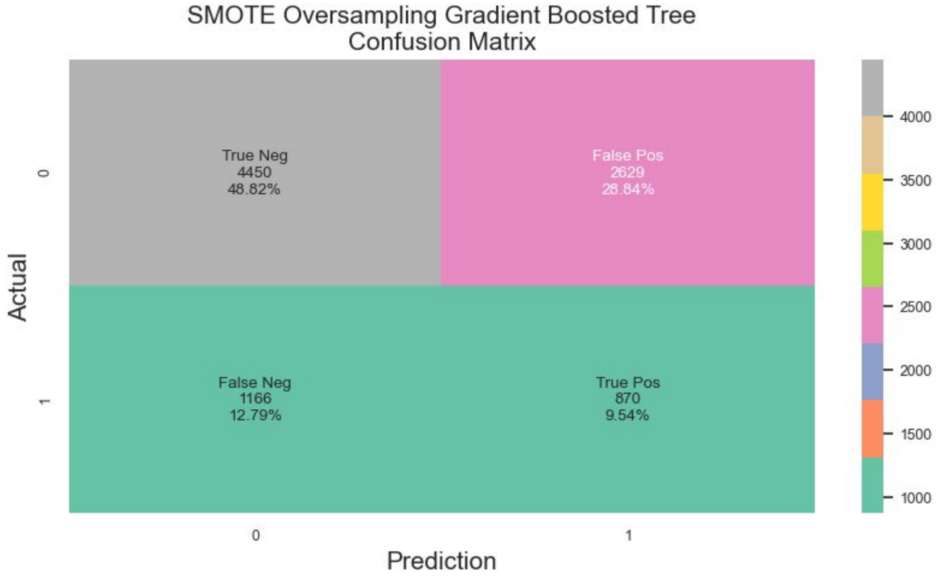
	Precision	Recall	f1-Score	Support
0	0.8591	0.8448	0.8519	7079
1	0.4898	0.5182	0.5036	2036
accuracy	0.7718	0.7718	0.7718	0
macro avg	0.6744	0.6815	0.6777	9115
weighted avg	0.7766	0.7718	0.7741	9115



SMOTE OVERSAMPLING
Gradient Boosted Tree

Classification Report

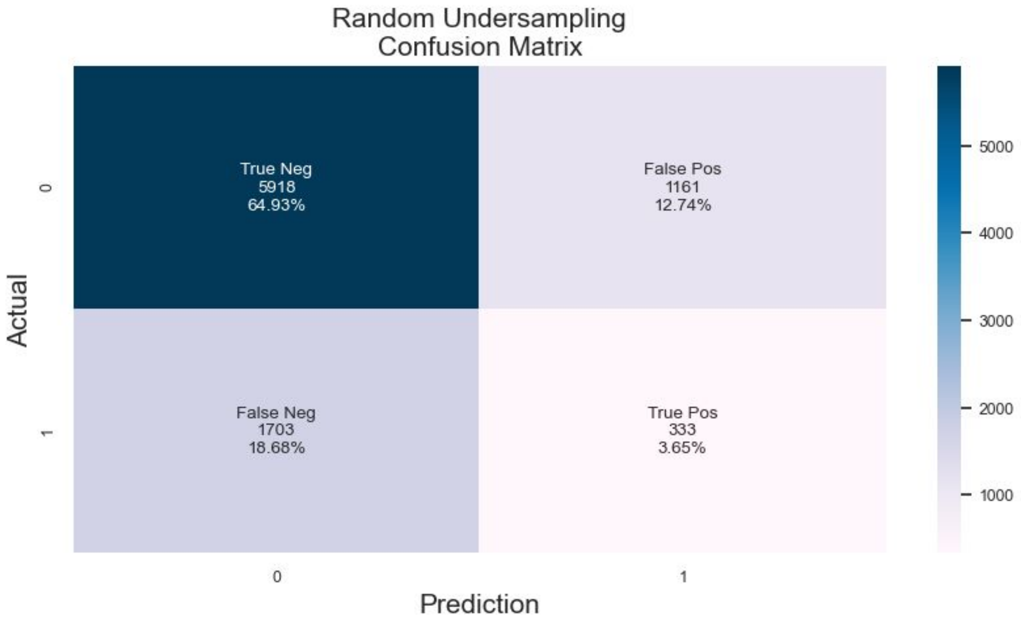
	Precision	Recall	f1-Score	Support
0	0.7924	0.6286	0.7011	7079
1	0.2486	0.4273	0.3144	2036
accuracy	0.5837	0.5837	0.5837	0
macro avg	0.5205	0.5280	0.5077	9115
weighted avg	0.6709	0.5837	0.6147	9115



RANDOM UNDERSAMPLING
Logistic Regression

Classification Report

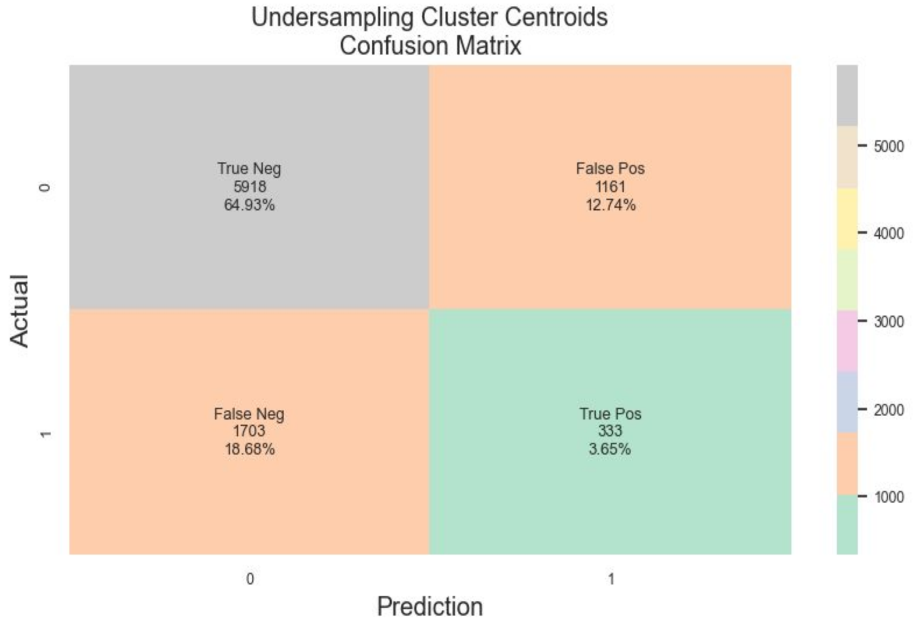
	Precision	Recall	f1-Score	Support
0	0.7924	0.6286	0.7011	7079
1	0.2486	0.4273	0.3144	2036
accuracy	0.5837	0.5837	0.5837	0
macro avg	0.5205	0.5280	0.5077	9115
weighted avg	0.6709	0.5837	0.6147	9115



UNDERSAMPLING
Cluster Centroids

Classification Report

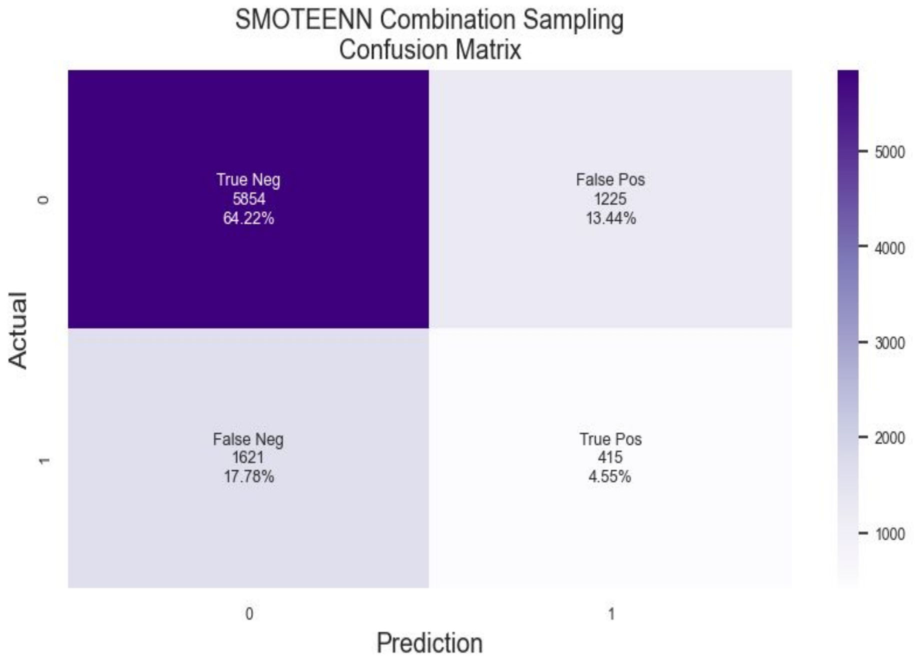
	Precision	Recall	f1-Score	Support
0	0.7924	0.6286	0.7011	7079
1	0.2486	0.4273	0.3144	2036
accuracy	0.5837	0.5837	0.5837	0
macro avg	0.5205	0.5280	0.5077	9115
weighted avg	0.6709	0.5837	0.6147	9115



COMBINATION SAMPLING SMOTEENN

Classification Report

	Precision	Recall	f1-Score	Support
0	0.7924	0.6286	0.7011	7079
1	0.2486	0.4273	0.3144	2036
accuracy	0.5837	0.5837	0.5837	0
macro avg	0.5205	0.5280	0.5077	9115
weighted avg	0.6709	0.5837	0.6147	9115

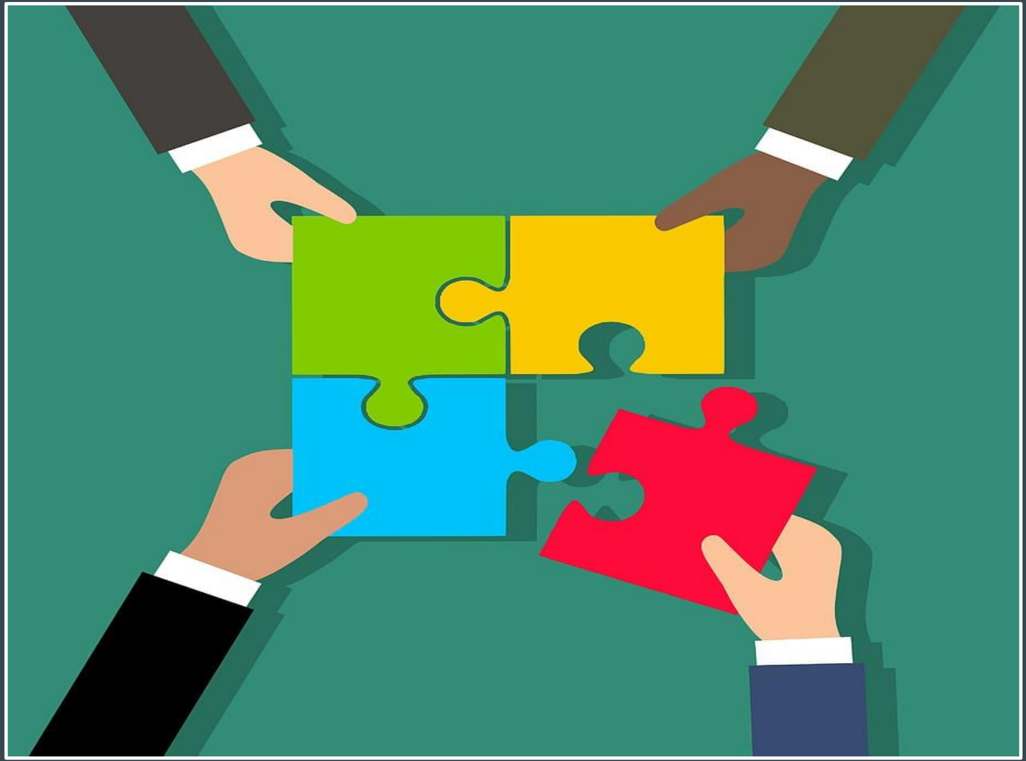


Dashboard Description

Tools: JavaScript, HTML

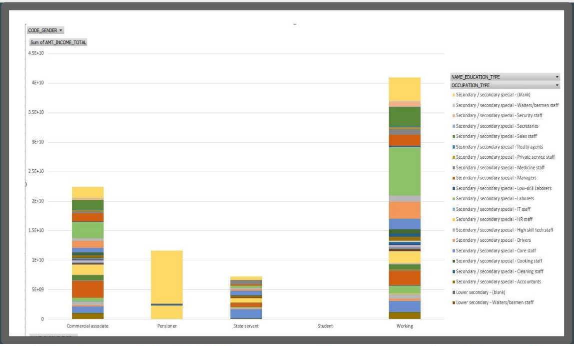
Interactive element(s)Features :

- Age
- Education
- Occupation
- Net income
- Number of family members



Credit Card Approval Prediction Dashboard

Using personal information and data submitted by credit card applicants,
the model will predict the probability of future defaults and credit card borrowings.



Approve or not?

The objective of this project is to help a financial institution to decide whether to issue a credit card to an applicant. Using personal information and data submitted by credit card applicants, the model will predict the probability of future defaults and credit card borrowings.

Filter Search

Enter Date

1/10/2010

Enter City

Toronto

Enter State

ON

Enter coutry

Canada

Enter a Shape

cicle

id [PK] character varying (10)	code_gender character varying (2)	flag_own_car character varying (2)	flag_own_realty character varying (2)	cnt_children integer	amt_income_total real	name_income_type character varying (40)	name_education_type character varying (40)	name_family_status character varying (40)
5008805	M	Y	Y	0	4.275	Working	Higher education	Civil marriage
5008806	M	Y	Y	0	1.125	Working	Secondary / secondary spec...	Married
5008808	F	N	Y	0	2.7	Commercial associate	Secondary / secondary spec...	Single / not married
5008809	F	N	Y	0	2.7	Commercial associate	Secondary / secondary spec...	Single / not married
5008810	F	N	Y	0	2.7	Commercial associate	Secondary / secondary spec...	Single / not married
5008811	F	N	Y	0	2.7	Commercial associate	Secondary / secondary spec...	Single / not married
5008812	F	N	Y	0	2.835	Pensioner	Higher education	Separated
5008813	F	N	Y	0	2.835	Pensioner	Higher education	Separated
5008814	F	N	Y	0	2.835	Pensioner	Higher education	Separated
5008815	M	Y	Y	0	2.7	Working	Higher education	Married
5112956	M	Y	Y	0	2.7	Working	Higher education	Married
5008819	M	Y	Y	0	1.35	Commercial associate	Secondary / secondary spec...	Married
5008820	M	Y	Y	0	1.35	Commercial associate	Secondary / secondary spec...	Married
5008821	M	Y	Y	0	1.35	Commercial associate	Secondary / secondary spec...	Married
5008822	M	Y	Y	0	1.35	Commercial associate	Secondary / secondary spec...	Married
5008823	M	Y	Y	0	1.35	Commercial associate	Secondary / secondary spec...	Married

Providence Bank
Credit Card Approval Model

Approve or not?

The objective of this project is to help a financial institution to decide whether to issue a credit card to an applicant. Using personal information and data submitted by credit card applicants, the model will predict the probability of future defaults and credit card borrowings.

Please enter your details
Age (Please enter in years)
Education
Please select
Occupation
Please select
Net Income (please enter annual salary)
Number of family members
Please select including yourself
Do you own property
Please Select

Age	Education	Occupation	Income	No. of Family Members	Own Realty	Approved 1 / Denied 0
33	Higher Education	No Occupation Type	$ 4,275	2	Y	1

WHAT WOULD WE DO DIFFERENTLY?

The Team

Binoy Luckoo

- Database schema
- Data cleaning & pre-processing
- Model Visualizations
- Google slides
- Dashboard

Samir Rifi

- Dataset sourcing
- Database
- Dashboard

Jane Huang

- Communications specialist
- Github Repository
- Google slides
- Dashboard
- AWS

Lucas Chandra

- GitHub Repository
- Database
- Data cleaning & pre-processing
- Machine learning models

QUESTIONS

CITATIONS

Slide 1 Background picture:
https://wowplus.net/these-are-the-new-upcoming-changes-to-your-credit-score-and-credit-cards/ (sept,2021)
Slide 4 pictures:
https://godmen.org/2021/02/20/best-credit-card-offers-what-are-the-best-offers/ (sept,2021)
Data:
https://www.kaggle.com/rikdifos/credit-card-approval-prediction