Presentation Content - *** This is just a placeholder - will be removed for final presentation ***

The presentation tells a cohesive story about the project and includes the following:

- Selected topic (Jane)
- Reason the topic is selected(Jane)
- Description of the source of data(Samir)
- Questions the team hopes to answer with the data(Samir)
- Description of the data exploration phase of the project (Lucas)
- Description of the analysis phase of the project (Lucas)
- Technologies, languages, tools, and algorithms used throughout the project (Binoy)
- Result of the analysis (Binoy)
- Recommendation for future analysis (Jane)
- Anything the team would have done differently (all four team members)

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.

Questions to be Answered by the Analysis & Models

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

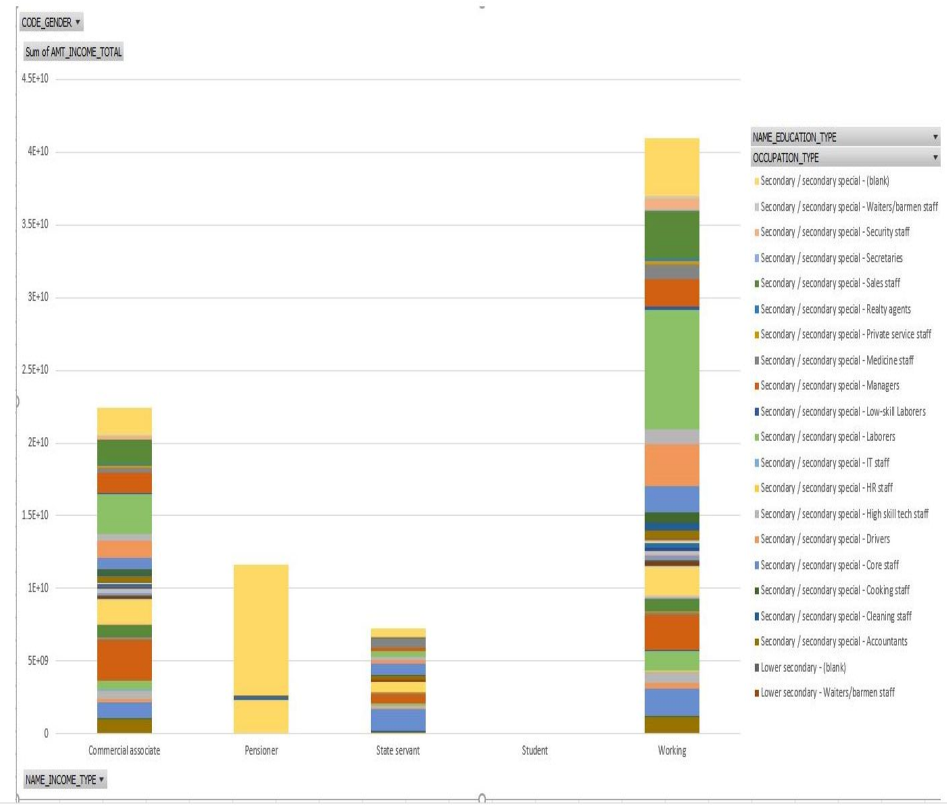
Database, Data Wrangling & Feature Engineering

Machine Learning Pipeline

Dashboard Presentation

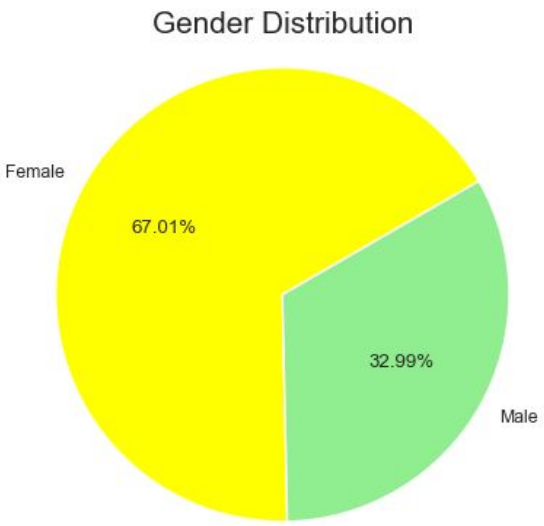
DataSweeper_Project

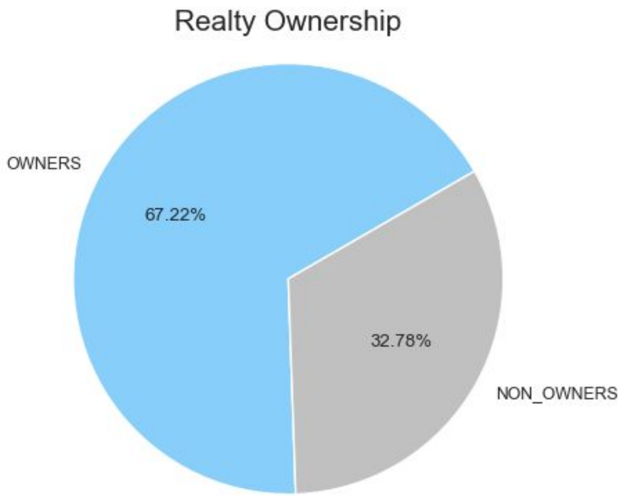
Data Exploration



DATASET DEMOGRAPHICS
Gender Distribution & Realty Ownership

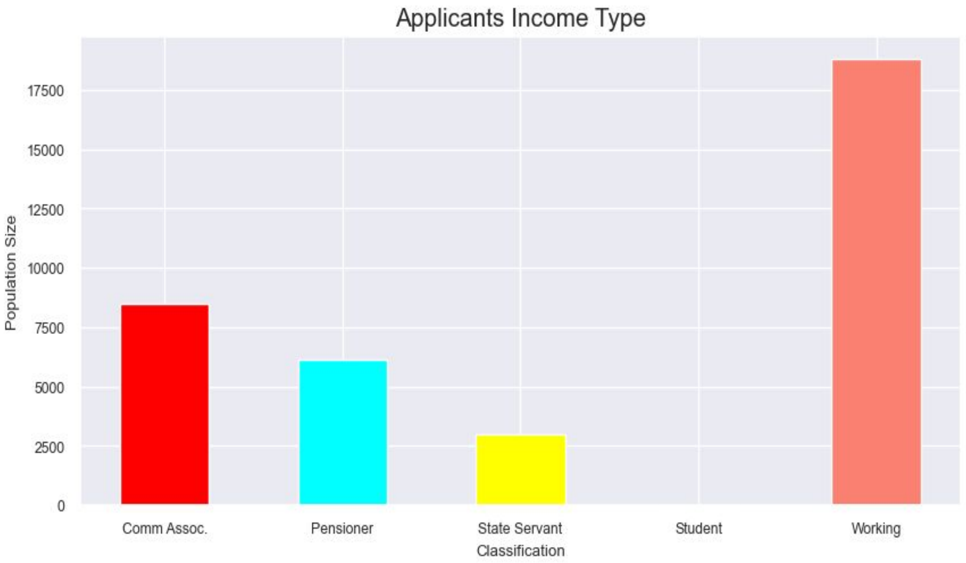
These charts show the gender distribution and realty ownership status of all applicants in the datasets being used for the models.

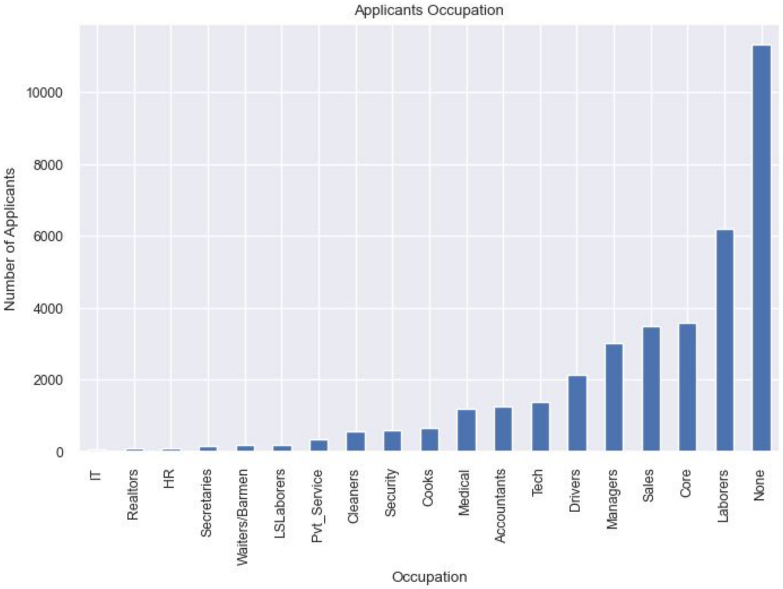




DATASET DEMOGRAPHICS
Applicants Income Type & Occupation

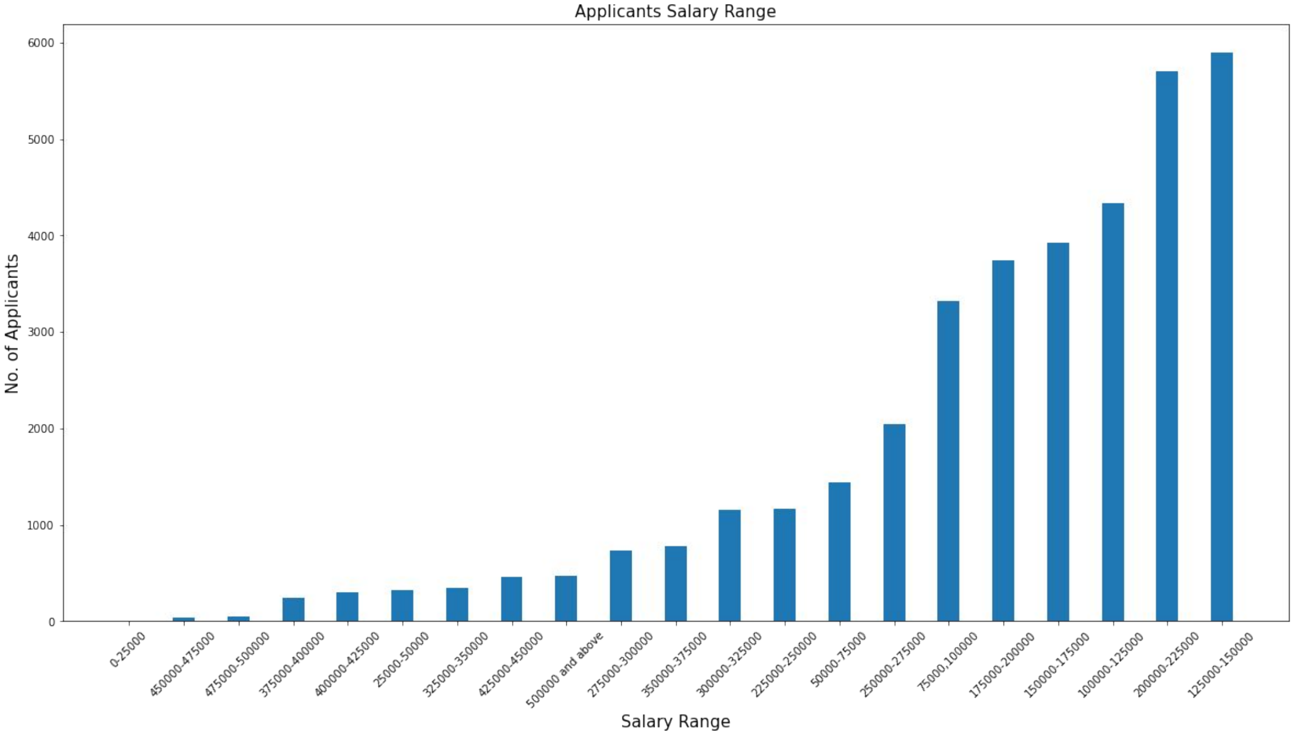
Applicants income type and occupation are displayed in the following charts





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	
Training & Testing Sets	Y value → X value →
Model Choice	Random Forest
Accuracy Scores	Training → Testing →

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”

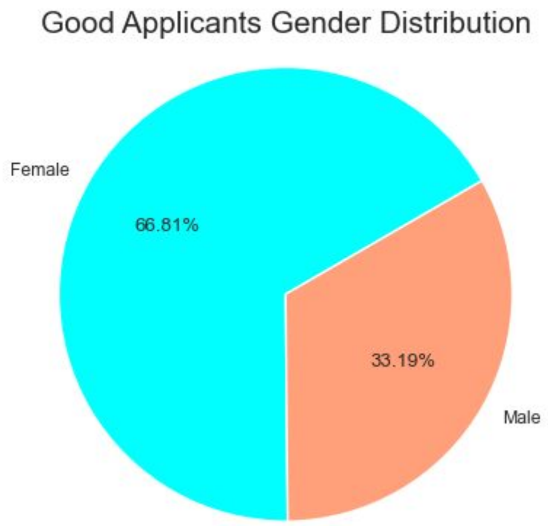
The following charts is an illustration of the above metrics for each model.

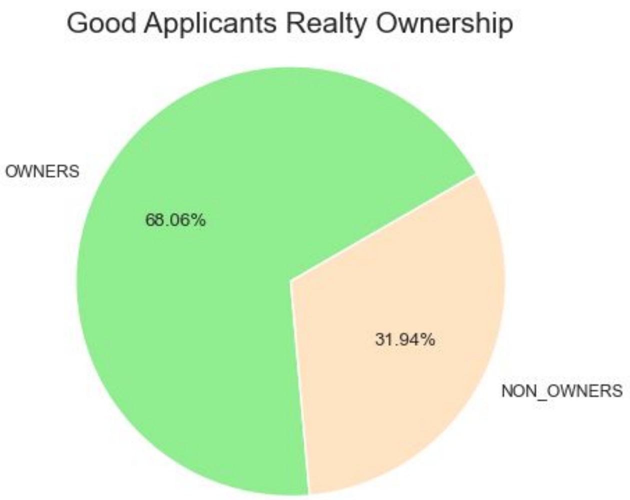
Data Analysis

To be reviewed

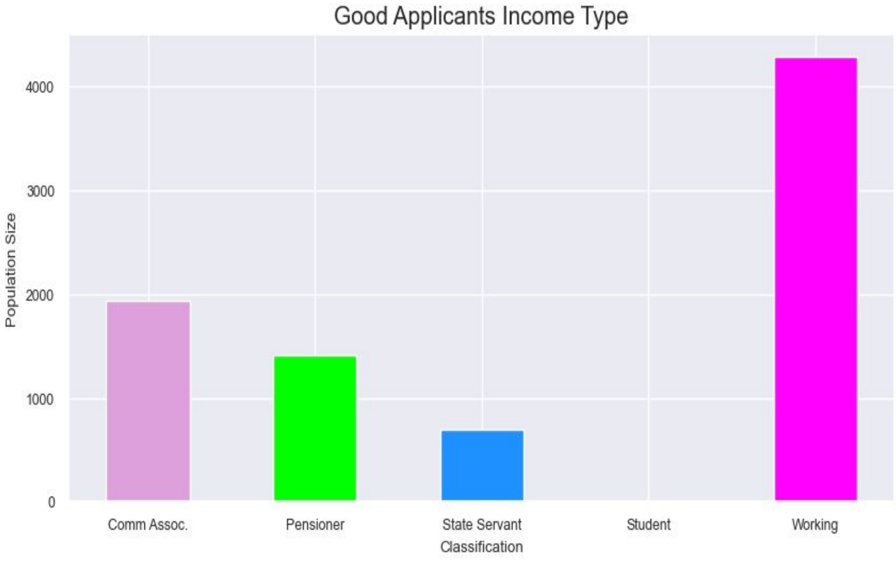
GOOD APPLICANTS DEMOGRAPHICS
Gender Distribution & Realty Ownership

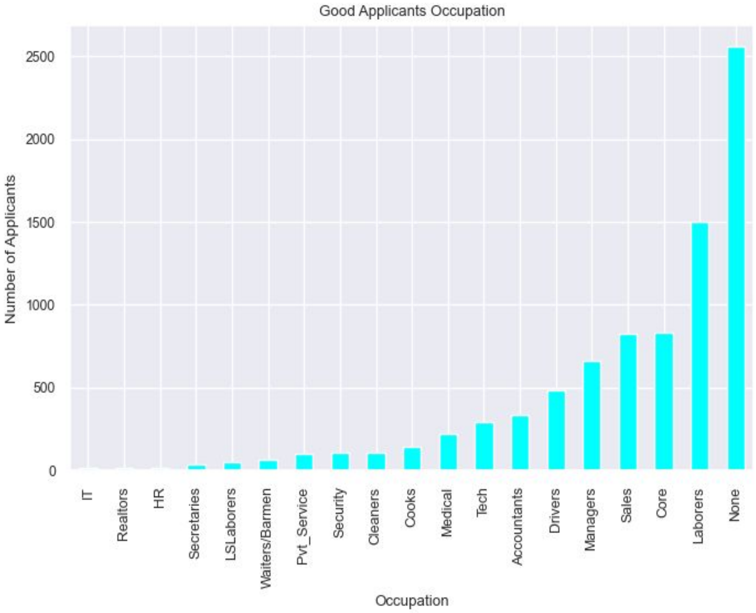
An analysis of the “good applicants” show that the distribution follows the same demographics as the whole population. This is depicted in the following charts.





GOOD APPLICANTS DEMOGRAPHICS
Income Type & Occupation





DATABASE

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

Oversampling - Summary of Results

Random Oversampling									
		""Good" Applicants			"Bad" Applicants				
Machine Learning Model	Accuracy Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Type I Error (FP)	Type II Error (FN)
Logistic Regression	0.51	0.23	0.50	0.32	0.78	0.51	0.62	37.87%	11.08%
Decision Tree	0.72	0.42	0.63	0.50	0.88	0.75	0.81	19.70%	8.21%
Random Forest	**0.73**	**0.42**	**0.61**	**0.50**	**0.87**	**0.76**	**0.81**	**18.74%**	**8.63%**
Gradient Boosted Tree	0.55	0.26	0.52	0.34	0.80	0.56	0.66	34.02%	10.67%
SMOTE Oversampling									
Logistic Regression	0.50	0.23	0.51	0.31	0.78	0.50	0.61	38.50%	11.04%
Decision Tree	0.77	0.48	0.51	0.49	0.86	0.84	0.85	12.44%	10.97%
Random Forest	0.77	0.49	0.52	0.51	0.86	0.84	0.85	**12.06%**	**10.71%**
Gradient Boosted Tree	0.62	0.26	0.37	0.30	0.79	0.69	0.74	24.22%	14.02%

Undersampling - Summary of Results

Rsandom Undersampling									
		""Good" Applicants			"Bad" Applicants				
Machine Learning Model	Accuracy Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Type I Error (FP)	Type II Error (FN)
Logistic Regression	0.62	0.26	0.37	0.30	0.79	0.69	0.74	37.62%	11.20%
Decision Tree	0.67	0.36	0.64	0.46	0.87	0.68	0.76	25.07%	8.13%
Random Forest	0.68	0.37	0.65	0.47	0.87	0.68	0.77	24.70%	7.76%

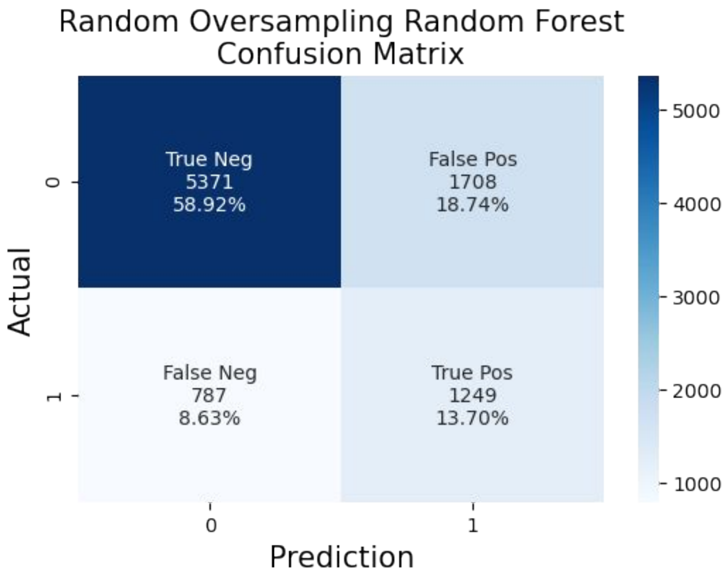
Combination Sampling - Summary of Results

Combination sampling - SMOTEENN									
		""Good" Applicants			"Bad" Applicants				
Machine Learning Model	Accuracy Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Type I Error (FP)	Type II Error (FN)
Logistic Regression	0.68	0.37	0.65	0.47	0.87	0.68	0.77	13.92%	17.97%

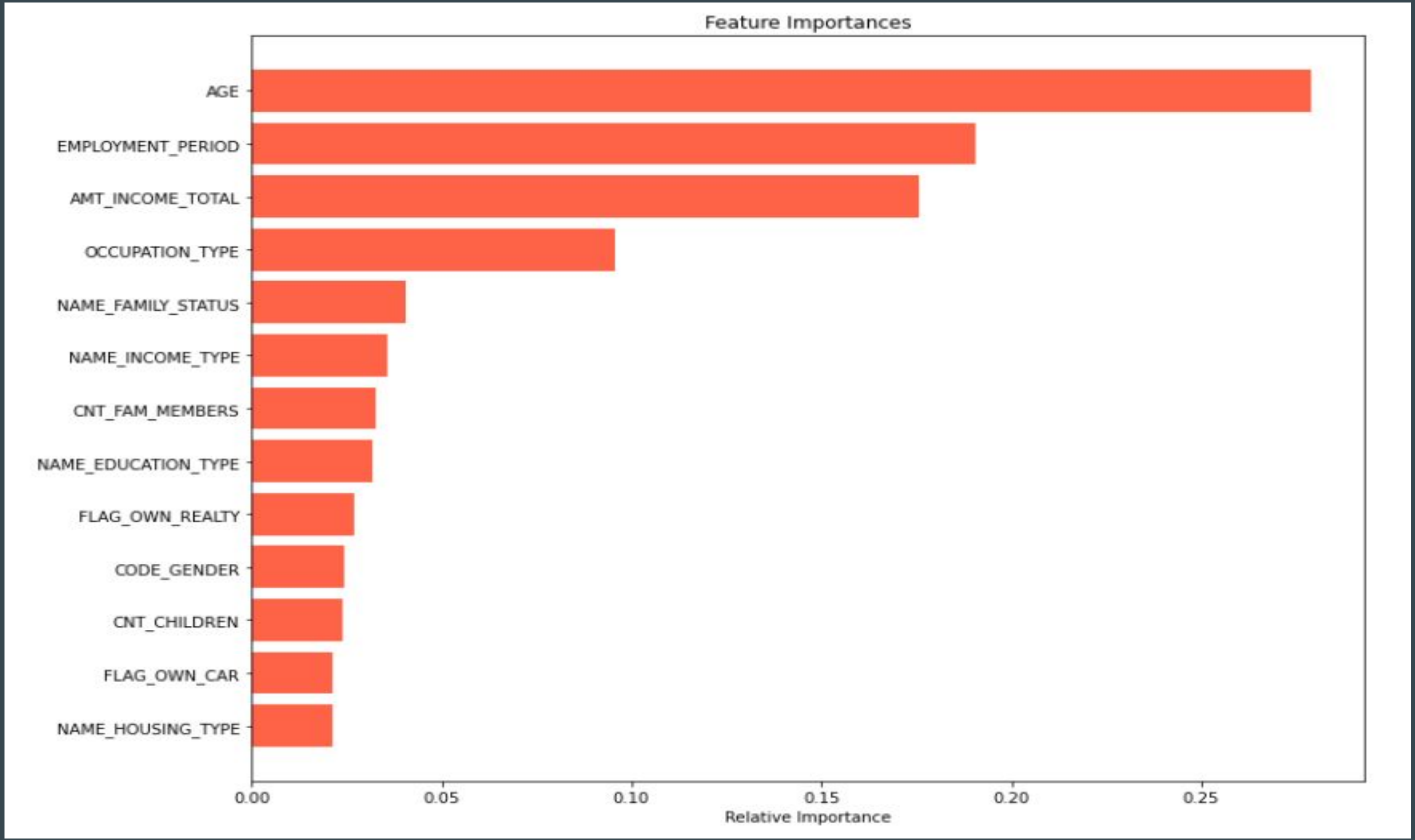
RANDOM OVERSAMPLING
Random Forest

Classification Report

	Precision	Recall	f1-Score	Support
0	0.87	0.76	0.81	7079
1	0.42	0.61	0.50	2036
accuracy	0.73	0.73	0.73	0
macro avg	0.65	0.69	0.66	9115
weighted avg	0.77	0.73	0.74	9115



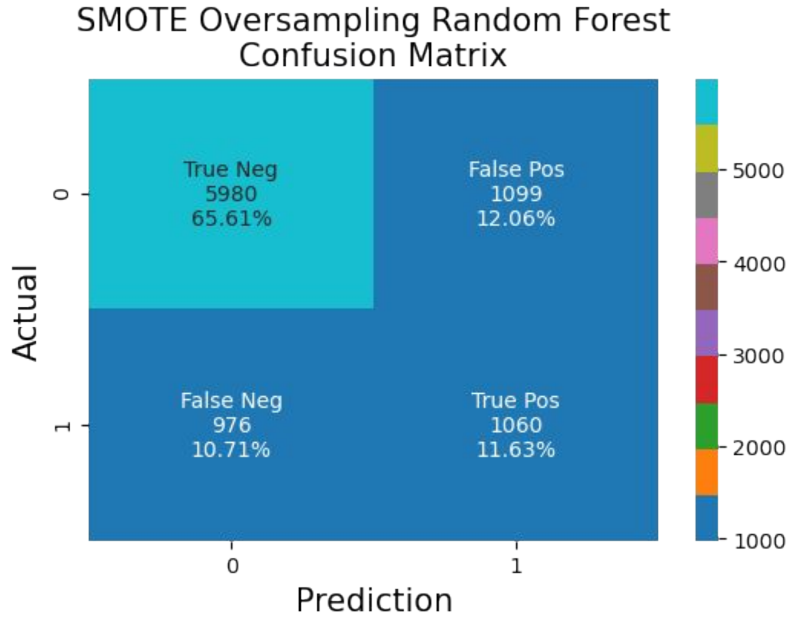
RANDOM OVERSAMPLING
Random Forest - Feature Importance



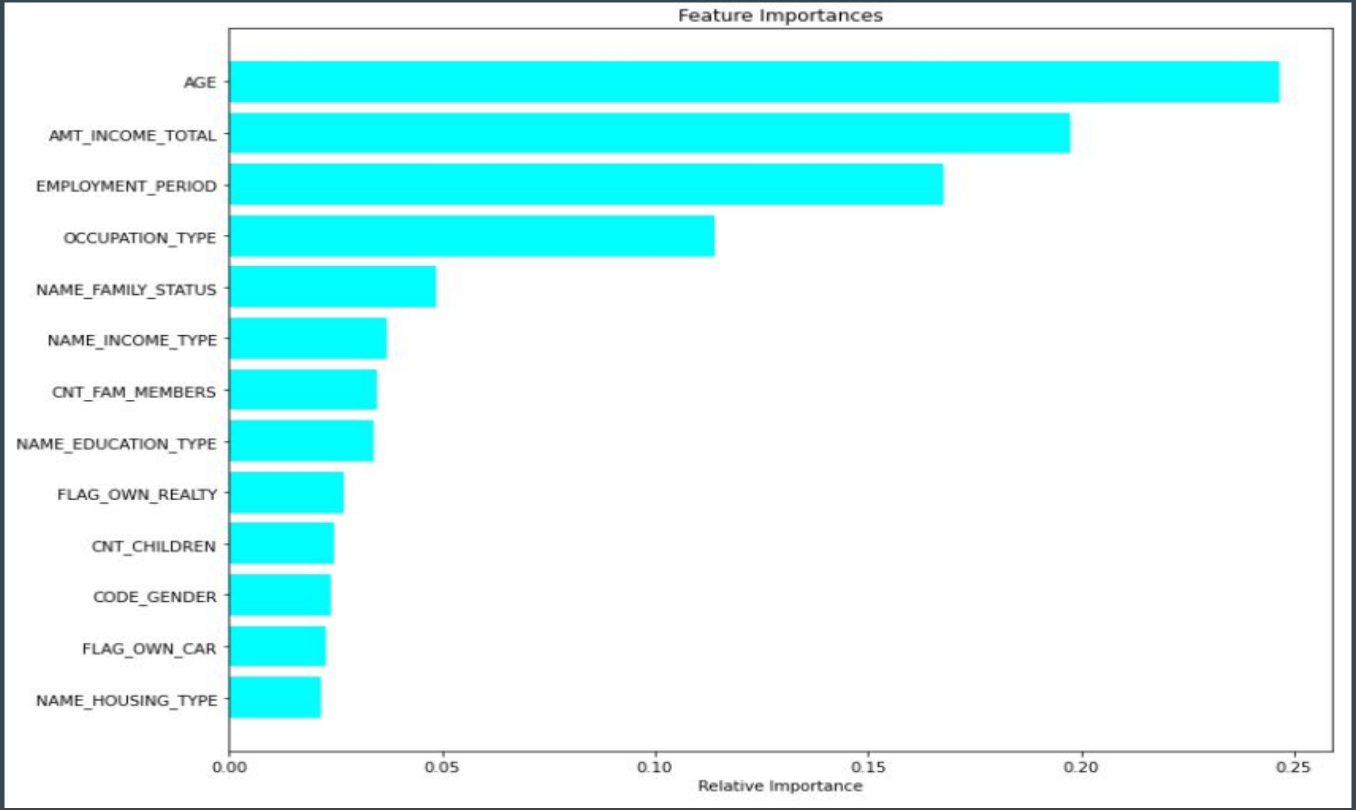
SMOTE OVERSAMPLING
Random Forest

Classification Report

	Precision	Recall	f1-Score	Support
0	0.86	0.84	0.85	7079
1	0.49	0.52	0.51	2036
accuracy	0.77	0.77	0.77	0
macro avg	0.68	0.68	0.68	9115
weighted avg	0.78	0.77	0.77	9115



SMOTE OVERSAMPLING
Random Forest - Feature Importance

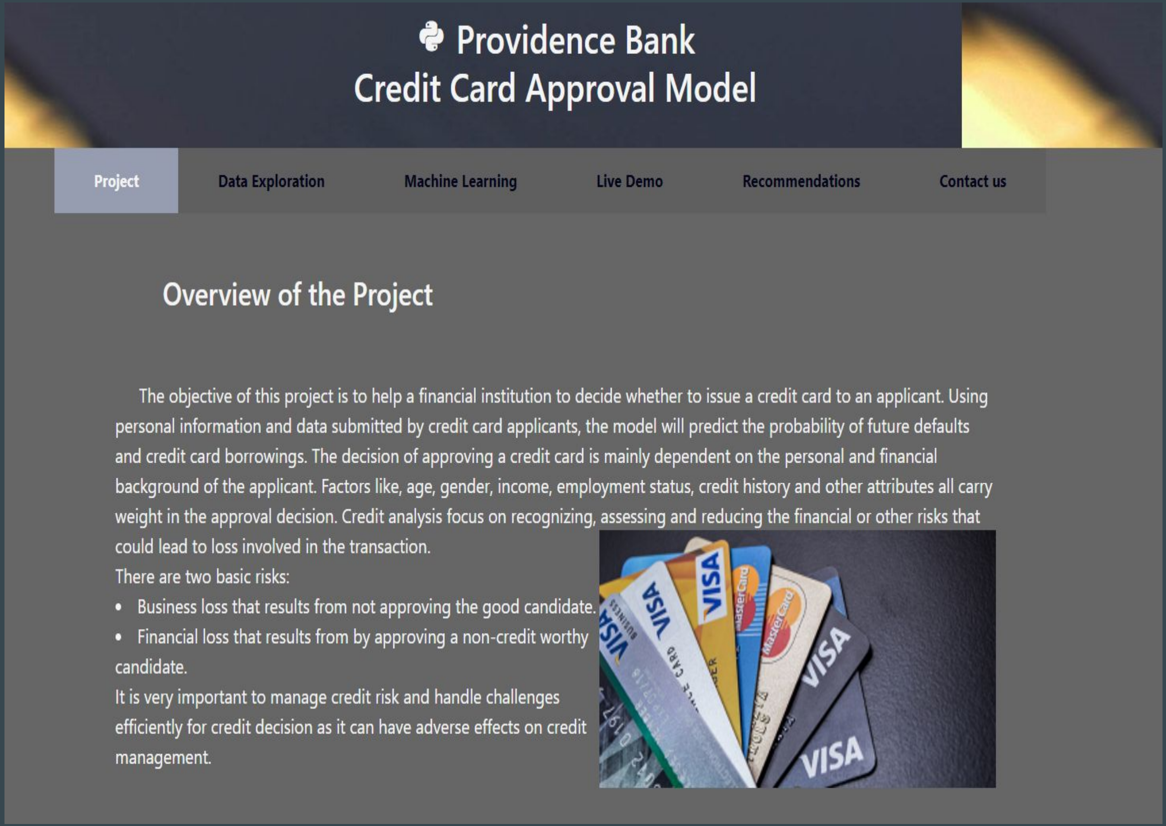


Dashboard Description

Tools: JavaScript, HTML

Interactive element(s)Features :

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



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

Cash card image https://www.nyra.com/aqueduct/racing/cash-card