Credit Risk Machine Learning

Project four by Zhao Wen, Rachel Woodill, and Maha Salman Cheema

Table of contents

01

Our Dataset

Dataset: Credit Card Fraud Detection

Repurposed our dataset from Project 1

After searching through a variety of different credit datasets, we decided to repurpose the same dataset we analyzed for Project 1.

- Obtained from Kaggle.
- Large dataset with over 200,000 rows and 122 columns pertaining to if an individual is a risk of being a credit defaulter.

Some of the relevant columns

columns_description

	Table	Row	Description
1	application_data	SK_ID_CURR	ID of loan in our sample
2	application_data	TARGET	Target variable (1 - client with payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample, 0 - all other cases)
5	application_data	NAME_CONTRACT_TYPE	Identification if loan is cash or revolving
6	application_data	CODE_GENDER	Gender of the client
7	application_data	FLAG_OWN_CAR	Flag if the client owns a car
8	application_data	FLAG_OWN_REALTY	Flag if client owns a house or flat
9	application_data	CNT_CHILDREN	Number of children the client has
10	application_data	AMT_INCOME_TOTAL	Income of the client
11	application_data	AMT_CREDIT	Credit amount of the loan
12	application_data	AMT_ANNUITY	Loan annuity
13	application_data	AMT_GOODS_PRICE	For consumer loans it is the price of the goods for which the loan is given
14	application_data	NAME_TYPE_SUITE	Who was accompanying client when he was applying for the loan

Preprocessing

Dropped unnecessary columns

SK_ID_CURR - the identification column was dropped

Null values removed

We dropped rows where the column contained more than 100,000 nulls

Oversampling the data

Our dataset was imbalanced in favour of non-defaulters, so to balance the data we used oversampling

02

Our Model

Target: Target column (0/1)
Features: All other columns

Initial attempt: Neural network model

- 91.9% accuracy but only predicting 0 as the outcome
- Data imbalanced in favour of non-defaulters
- Over-sampling used to correct error
- Lead to 50% accuracy

Decision Tree Models

- A form of supervised learning
- Used to categorize or make predictions based on previous data
- Base is called the root node, from which the decision nodes flow

Final model results

	Predicted 0	Predicted 1
Actual 0	33989	4157
Actual 1	8	38609

Confusion matrix: significant number of false positives.
Predicting defaulting when there is none.

Accuracy Score : 0.9457420892878079
Classification Report

	precision	recall	f1-score	support
0	1.00	0.89	0.94	38146
1	0.90	1.00	0.95	38617
accuracy			0.95	76763
macro avg	0.95	0.95	0.95	76763
weighted avg	0.95	0.95	0.95	76763

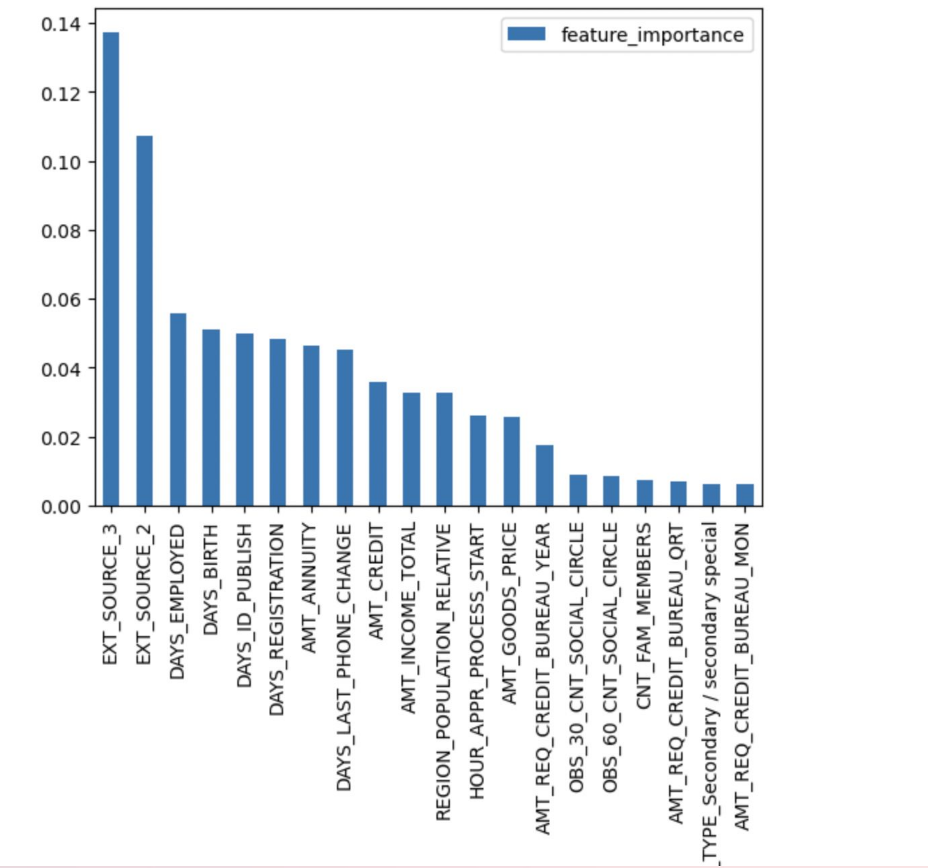
Most important features:

EXT_SOURCE_3 is described as a normalized score from external source.

EXT_SOURCE_2 is described as a normalized score from external source.

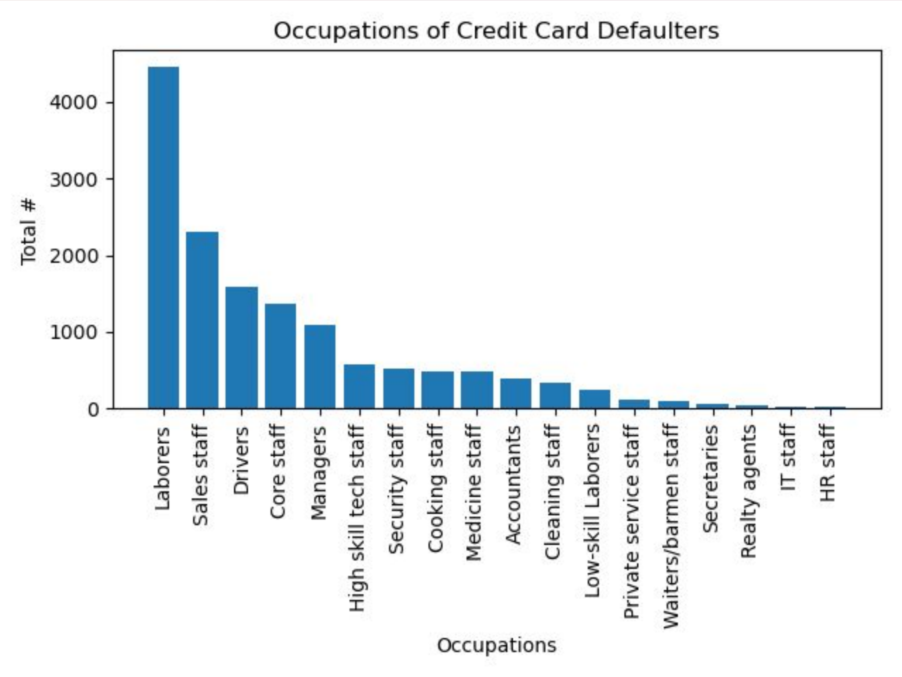
Both are measures of credit score.

DAYS_EMPLOYED is described as how many days the individual was employed before the application.

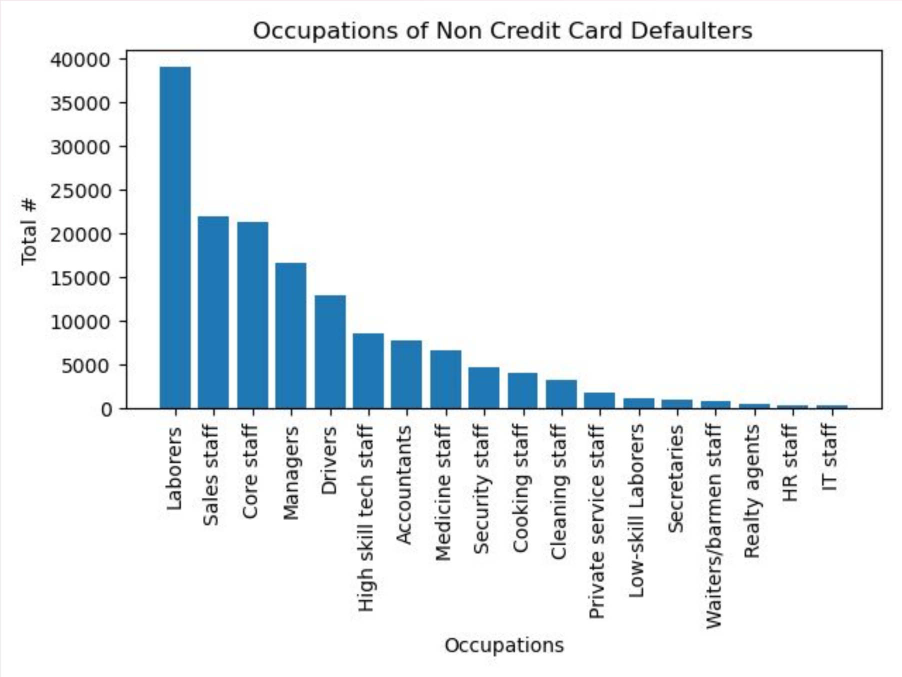


03

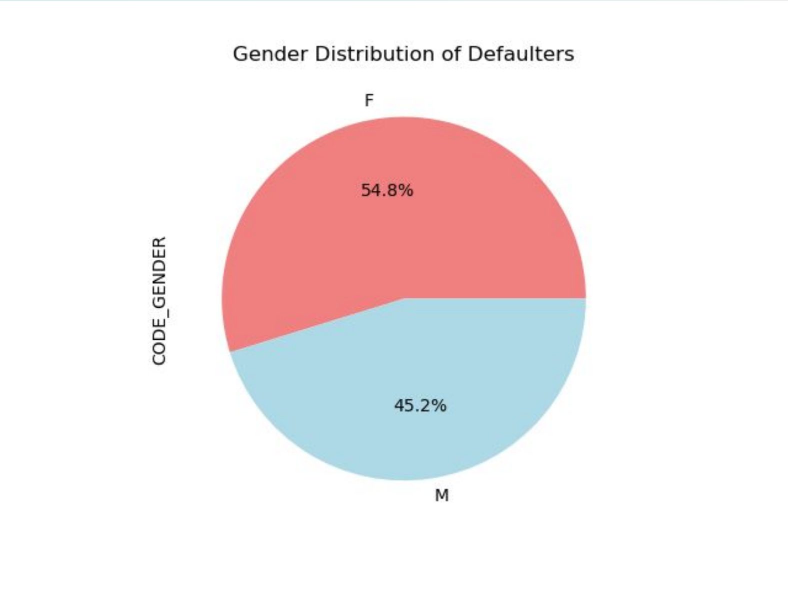
Further data exploration

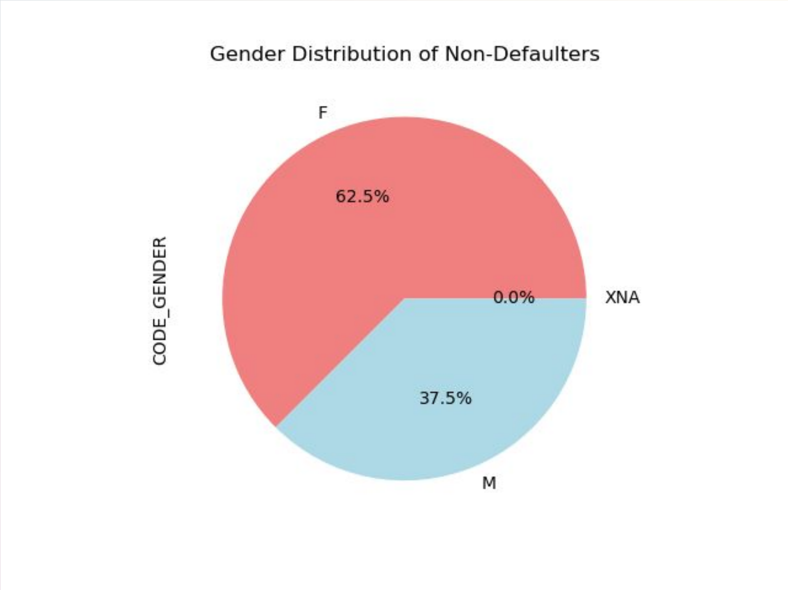


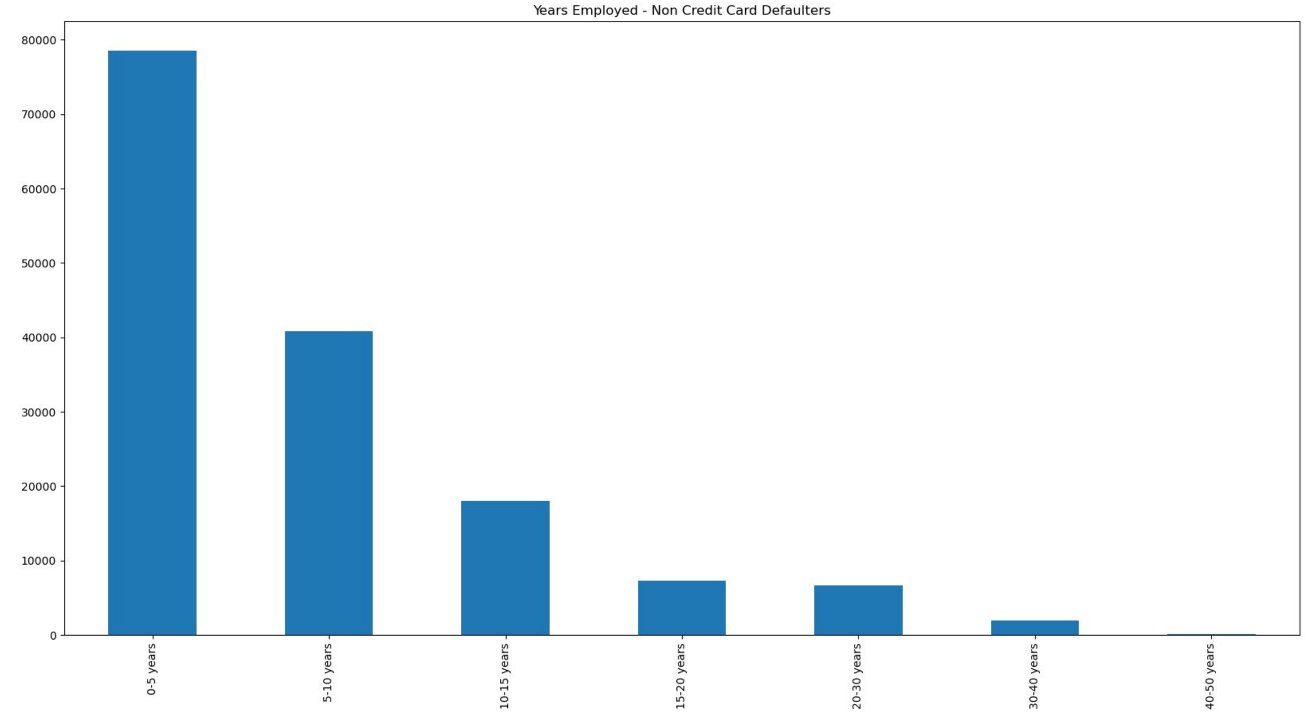
Occupations of Credit Card Defaulters
Total #
0
1000
2000
3000
4000
Occupations
Laborers
Sales staff
Drivers
Core staff
Managers
High skill tech staff
Security staff
Cooking staff
Medicine staff
Accountants
Cleaning staff
Low-skill Laborers
Private service staff
Waiters/barmen staff
Secretaries
Realty agents
IT staff
HR staff



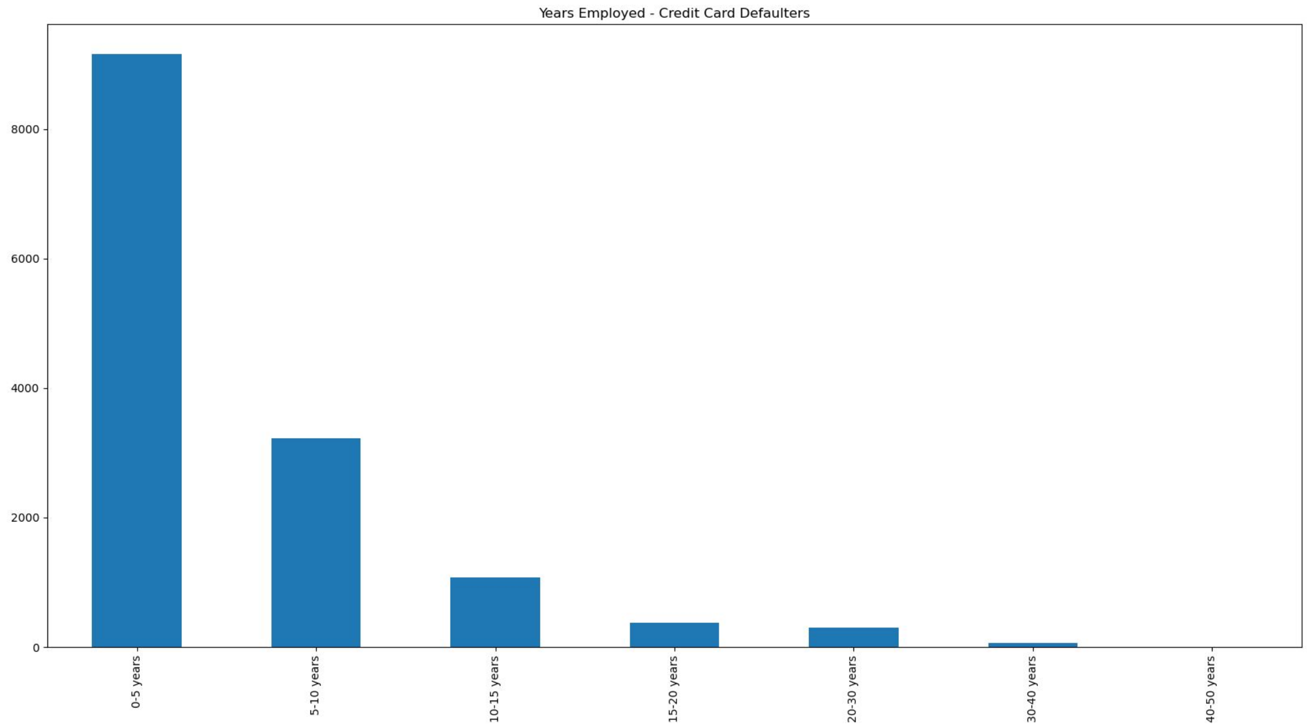
Occupations of Non Credit Card Defaulters
Total #
0
5000
10000
15000
20000
25000
30000
35000
40000
Occupations
Laborers
Sales staff
Core staff
Managers
Drivers
High skill tech staff
Accountants
Medicine staff
Security staff
Cooking staff
Cleaning staff
Private service staff
Low-skill Laborers
Secretaries
Waiters/barmen staff
Realty agents
HR staff
IT staff


Gender Distribution of Defaulters
CODE_GENDER
F
54.8%
45.2%
M


Gender Distribution of Non-Defaulters
CODE_GENDER
F
62.5%
0.0%
XNA
37.5%
M



Years Employed - Non Credit Card Defaulters
80000
70000
60000
50000
40000
30000
20000
10000
0
0-5 years
5-10 years
10-15 years
15-20 years
20-30 years
30-40 years
40-50 years


Years Employed - Credit Card Defaulters
8000
6000
4000
2000
0
0-5 years
5-10 years
10-15 years
15-20 years
20-30 years
30-40 years
40-50 years

04

Analysis

What does this mean?

From our model, it is possible to predict credit card defaulting using a Decision Tree classifier.

05

Conclusions

Challenges

- Our dataset contained a large portion of null values that made it difficult to predict defaulting, without cleaning up the columns.
- The target was very imbalanced
- There are many low-relevance features which can cause overfitting

Thanks!

Do you have any questions?