Credit Analysis Case Study

MPOWER Financing Summer 2020 Internship

Author: Sravan Sreeram

02/19/2020

Dallas, TX

Table of Contents

SUMMARY

This report involves the observations of Early Risk Score of college students, the analysis of factors affecting the Early Risk Score and conclusive business recommendations.

INTRODUCTION

The analysis done for this report was using two datasets, “Origination Data.csv” and “Performance.xlsx” that were given for this study.

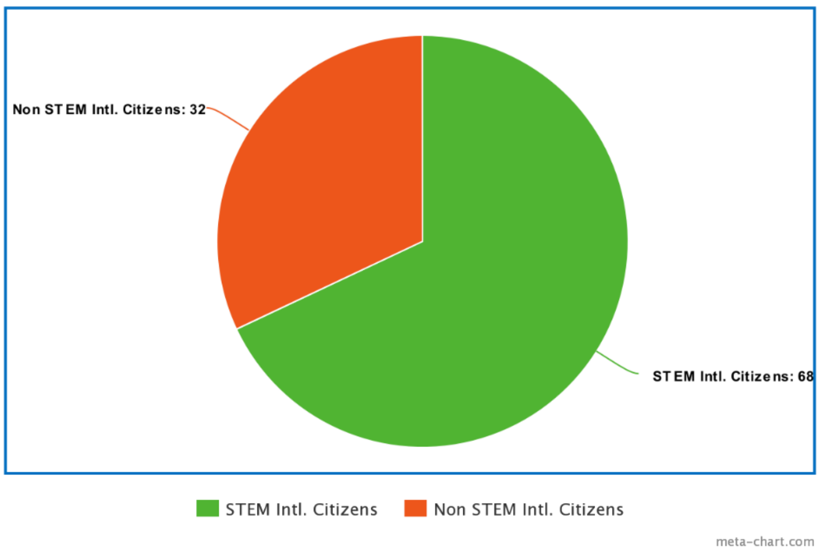
The main aim is to derive meaningful insights from the given data and identify the major influencers of the Early Risk Score.

An essential part of this process involved weeding out irregularities and unresourceful features in the data; thereby creating a unified, trim and workable dataset.

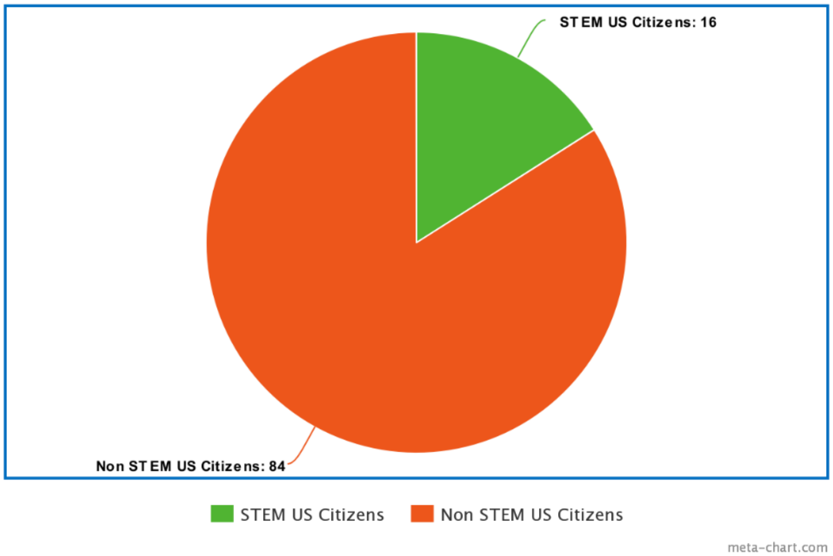
ANALYSIS

1. Initial observation using Excel functions

We begin the analysis by setting Early Risk Score as our target variable and exploring the data using MS Excel and try to derive a sense of the data by performing filter operations and creating pivot tables.



From the above pie chart, it is observed that among the students who are classified as International (non-US) citizens, 68% are enrolled in STEM courses and only around 32% are enrolled in non-STEM courses



From the above pie chart, it is observed that among the students who are classified as - US citizens, only 16% are enrolled in STEM courses and about 84% are enrolled in non-STEM courses

- On average, students with SSN have a lower Early Risk Score (0.44) when compared to students without SSN (0.56)
- On average, students enrolled in STEM programs have a lower Early Risk Score (0.35) when compared to students enrolled in non-STEM programs (0.53)
- The average Early Risk Score of non-US citizen students enrolled in STEM programs and US citizen students enrolled in STEM programs is around the same at 0.3

2. Creating a unified dataset

We were provided with two files, "Origination Data.csv" and "Performance.xlsx". Both these files had one column common between them, which is Loan Number.

Before merging files, the duplicate entries on the excel files were removed. The resultant Database consisted of single entries of Loan number and all duplicates and null values in this column were removed.

Loan Number	Tell Us Abo	Has SSN	US Citizen	Enrollment Status	STEM	Credit Scor	Credit Scor	GPA	Approved Loan Amount	Interest Ra	Test Loan	Early Risk Score
35		Yes	FALSE	2nd Year Graduate	No	630	0	3.74	15500	11.99	0	0
43		Yes	FALSE	2nd Year Graduate	No	644		3.86	10446	11.99	0	0
63		Yes	FALSE	4th Year Undergraduate	No	688		3.2	15000	13.99	0	0
65		No	FALSE	1st Year Graduate	No			0	7000	11.99	0	0.5
66		No	TRUE	3rd Year Undergraduate	No	641		3	15000	9.99	0	0
67		No	FALSE	2nd Year Graduate	No	656		3.33	15000	11.99	0	0
69		Yes	FALSE	2nd Year Graduate	No		757	3.3	8000	11.99	0	0
70		Yes	FALSE	1st Year Graduate	No	644		3.88	20000	11.99	0	0
71		Yes	FALSE	3rd Year Graduate	No	670		3.52	25000	11.99	0	0
74		Yes	FALSE	2nd Year Graduate	No	682		3.23	23000	11.99	0	0
83		Yes	FALSE	2nd Year Graduate	No	614		3.33	6000	11.99	0	1.4

3. Cleaning the unified dataset

The newly created dataset is imported into Python Jupyter Notebook for further cleaning and analysis

For simplification of the task, we drop "Tell Us About You" and "Loan Number" from our further analysis

```
LoanData = pd.read_csv(data_path)
```

```
In [68]: LoanData = LoanData.drop(columns=['Loan Number', 'Tell Us About You')
         LoanData
```

Out[68]:

	Has SSN	US Citizen	Enrollment Status	STEM	Credit Score 1	Credit Score 2	GPA	Approved Loan Amount	Interest Rate	Test Loan	Early Risk Score
0	Yes	False	2nd Year Graduate	No	630.0	0.0	3.74	15500.0	11.99	0	0.0
1	Yes	True	4th Year Undergraduate	No	536.0	NaN	3.20	10000.0	9.99	0	NaN
2	Yes	False	2nd Year Graduate	No	644.0	NaN	3.86	10446.0	11.99	0	0.0
3	Yes	False	4th Year Undergraduate	No	688.0	NaN	3.20	15000.0	13.99	0	0.0
4	No	False	1st Year Graduate	No	NaN	NaN	0.00	7000.0	11.99	0	0.5
...	...	...	...	...	...	...	...	...	...	...	...
1024	Yes	False	3rd Year Graduate	No	606.0	0.0	3.90	10500.0	11.99	0	0.0
1025	Yes	False	2nd Year Graduate	No	764.0	0.0	3.78	8000.0	11.99	0	0.0
1026	No	False	2nd Year Graduate	No	687.0	NaN	0.00	25000.0	11.99	0	0.0
1027	No	False	1st Year Graduate	No	NaN	0.0	0.00	4000.0	11.99	0	0.0
1028	No	False	1st Year Graduate	Yes	NaN	0.0	0.00	23000.0	11.99	0	0.0

1029 rows × 11 columns

We then look for information about missing values in the dataset

```
# information about missing values in dataset
LoanData.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1029 entries, 0 to 1028
Data columns (total 11 columns):
Has SSN                 1029 non-null object
US Citizen              1029 non-null bool
Enrollment Status       1029 non-null object
STEM                    986 non-null object
Credit Score 1          469 non-null float64
Credit Score 2          684 non-null float64
GPA                     1029 non-null float64
Approved Loan Amount    1029 non-null float64
Interest Rate           1029 non-null float64
Test Loan               1029 non-null int64
Early Risk Score        976 non-null float64
dtypes: bool(1), float64(6), int64(1), object(3)
memory usage: 81.5+ KB
```

In order to have a standardized dataset, we impute the missing values numerical features with the median value. In this case, since many values in the Credit Score 1 and Credit Score 2 columns are empty, we replace the empty cells with the corresponding median values.

4. Correlation Matrix and Heatmap analysis of numerical variables

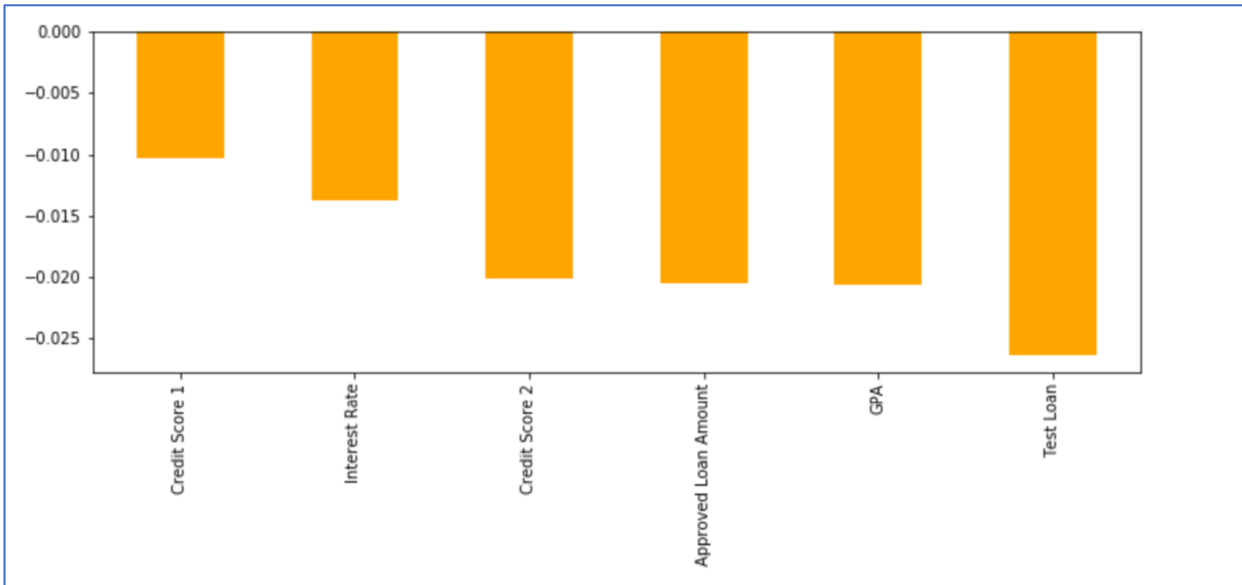
In order to find the most influential feature that explains the Early Risk Score, we create a correlation matrix and heatmap with the numerical features in the dataset

```
#heatmap

corr = numerical_features_df.corr()
corr.style.background_gradient(cmap='coolwarm')
```

	Credit Score 1	Credit Score 2	GPA	Approved Loan Amount	Interest Rate	Test Loan	Early Risk Score
Credit Score 1	1	-0.0444432	-0.093599	0.179089	0.0501465	-0.0345039	-0.0103372
Credit Score 2	-0.0444432	1	0.0513704	-0.0367086	-0.0343707	-0.0312368	-0.0200967
GPA	-0.093599	0.0513704	1	-0.233784	0.12681	-0.0176386	-0.0206063
Approved Loan Amount	0.179089	-0.0367086	-0.233784	1	0.0623103	0.0502355	-0.0204821
Interest Rate	0.0501465	-0.0343707	0.12681	0.0623103	1	-0.0649136	-0.0137263
Test Loan	-0.0345039	-0.0312368	-0.0176386	0.0502355	-0.0649136	1	-0.0264083
Early Risk Score	-0.0103372	-0.0200967	-0.0206063	-0.0204821	-0.0137263	-0.0264083	1

Based on the heatmap, we arrive at the below bar plot of the highly correlated variables with the target variable, Early Risk Score.



From the bar plot, we can observe that among the numerical features in the dataset, the highly correlated features with Early Risk Score are Test Loan, GPA and Approved Loan Amount. These features are inversely proportional to the Early Risk Score.

- It can be inferred that; higher student GPA might result in a lower Early Risk Score.
- It can also be inferred that; students with low Early Risk Score have higher approved loan amounts.

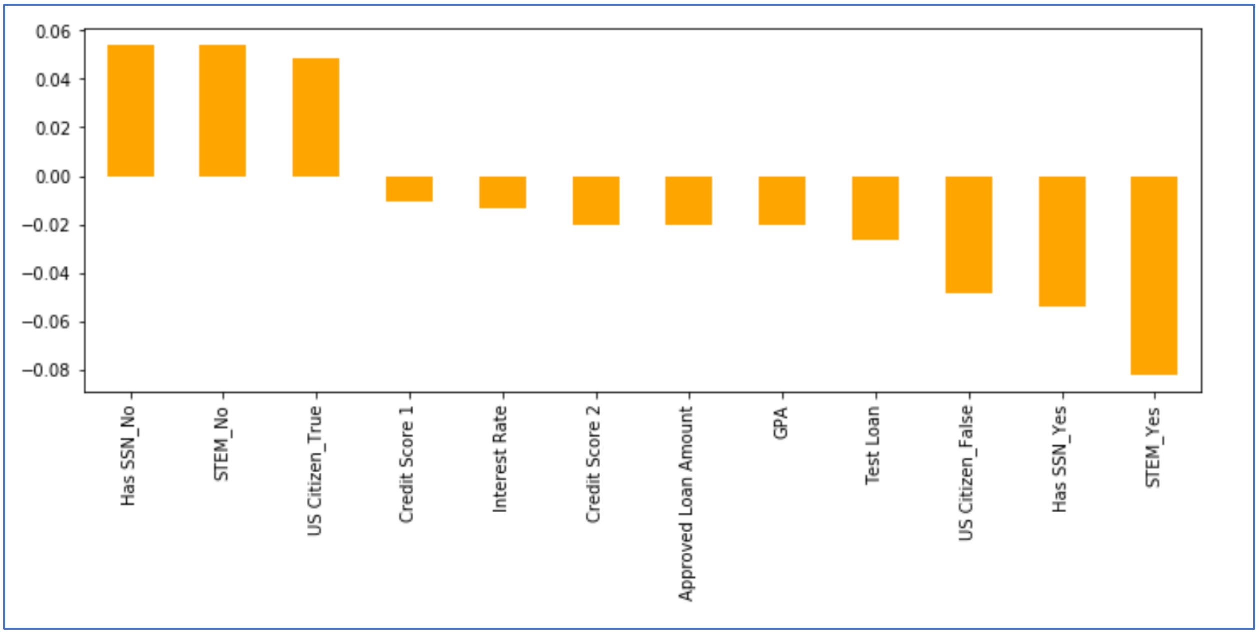
5. One-Hot Encoding of data & Heatmap to derive further insights

Thus far, the analysis was done primarily based on numerical features in the dataset. But, many important categorical features such as "STEM", "US Citizen" and "Has SSN" were left out. Therefore, we perform One-Hot encoding on such variables and convert them into numerical features.

After this conversion, we again perform the correlation matrix and heatmap to observe changes in the correlated variables.

	Credit Score 1	Credit Score 2	GPA	Approved Loan Amount	Interest Rate	Test Loan	Early Risk Score	US Citizen_False	US Citizen_True	STEM_No	STEM_Yes
Credit Score 1	1	-0.0444432	-0.093599	0.179089	0.0501465	-0.0345039	-0.0103372	0.105401	-0.105401	-0.000579731	-0.0174675
Credit Score 2	-0.0444432	1	0.0513704	-0.0367086	-0.0343707	-0.0312368	-0.0200967	0.0870492	-0.0870492	0.00407906	0.0225933
GPA	-0.093599	0.0513704	1	-0.233784	0.12681	-0.0176386	-0.0206063	-0.0602791	0.0602791	-0.0662417	0.0674946
Approved Loan Amount	0.179089	-0.0367086	-0.233784	1	0.0623103	0.0502355	-0.0204821	0.172621	-0.172621	-0.0611171	0.0658445
Interest Rate	0.0501465	-0.0343707	0.12681	0.0623103	1	-0.0649136	-0.0137263	0.579768	-0.579768	-0.144618	0.17727
Test Loan	-0.0345039	-0.0312368	-0.0176386	0.0502355	-0.0649136	1	-0.0264083	-0.0385481	0.0385481	-0.000633393	0.00772748
Early Risk Score	-0.0103372	-0.0200967	-0.0206063	-0.0204821	-0.0137263	-0.0264083	1	-0.0488049	0.0488049	0.0539638	-0.082121
US Citizen_False	0.105401	0.0870492	-0.0602791	0.172621	0.579768	-0.0385481	-0.0488049	1	-1	-0.0470609	0.0746499
US Citizen_True	-0.105401	-0.0870492	0.0602791	-0.172621	-0.579768	0.0385481	0.0488049	-1	1	0.0470609	-0.0746499
STEM_No	-0.000579731	0.00407906	-0.0662417	-0.0611171	-0.144618	-0.000633393	0.0539638	-0.0470609	0.0470609	1	-0.906606
STEM_Yes	-0.0174675	0.0225933	0.0674946	0.0658445	0.17727	0.00772748	-0.082121	0.0746499	-0.0746499	-0.906606	1
Has SSN_No	-0.101147	0.0592777	-0.289231	0.215779	-0.0132207	0.0147808	0.054107	0.0993464	-0.0993464	0.0605587	-0.0435254
Has SSN_Yes	0.101147	-0.0592777	0.289231	-0.215779	0.0132207	-0.0147808	-0.054107	-0.0993464	0.0993464	-0.0605587	0.0435254

Based on the heatmap, we arrive at the below bar plot of the highly correlated variables with the target variable, Early Risk Score.



From the new bar plot, it can be observed that the newly added features seem to be highly correlated with the Early Risk Score variable, replacing variables like Approved Loan Amount and GPA, that were earlier observed as highly correlated.

- It can be inferred for this bar plot that, enrolling in a STEM program, having SSN and being an international student suggests having lower Early Risk Score.
- It can also be inferred that, enrolling in a STEM program or not having an SSN can result in a higher Early Risk Score.

Conclusion & Recommendations

Based on the above analysis, we can make the following conclusions and recommendation:

1. Students enrolled in STEM generally have a lower Early Risk Score and hence, providing loan to such students involved lower risks and higher probability of loan repayment.

2. It seen that majority of the US citizen seem to enroll in non-STEM programs and since non-STEM programs are positively correlated with Early Risk Score, therefore the Early Risk Score of US citizen is higher when compared to international students.

3. It seen that majority of the non-US citizen seem to enroll in STEM programs and since STEM programs are negatively correlated with Early Risk Score, therefore the Early Risk Score of non-US citizen is lower when compared to US citizen students.

4. An important and interesting observations is that; while the Early Risk Score of non-US citizen students are lower, the Interest rates and approved loan amounts are higher compared to US citizen students.

   Based on the above research and analysis it can therefore recommend that, providing loan to non-US citizen students enrolling in STEM programs will be the most profitable as these students generally have low risks associated as Early Risk Scores are low, overwhelmingly opt for STEM designated programs and accept the loan at a higher interest rate. Moreover, the approved loan amount is also higher.

Resources

- MS Excel
- Jupyter Notebooks & Python libraries (NumPy, pandas, sci-kit learn)
- Meta-chart.com
- Origination Data.csv
- Performance.xlsx