Home Loan Default Prediction on the basis of Income Rates

DEPARTMENT OF INFORMATION SYSTEMS & DECISION SCIENCES

UNIVERSITY OF SOUTH FLORIDA

ISM 6136: DATA MINING

GROUP PROJECT

Project Contributors:

Ankit Singh
Vignesh Ashok Kumar
Vinay Murthy
Yaassh Rao

Abstract

In the world of finance, Credit or Default Risk is the probability that companies or individuals will be unable to make the required payments on their debt obligations. Lenders are exposed to default risk on every extension of credit.

Default risk is comprised of two components:

Systemic Risk - The risk related to an external system of interconnected relationships (think of this as the risk associated with they economy, interest rate changes, inflation, recessions, and wars). This risk is often considered unavoidable. The only option is to hedge or not be in the market.

Unsystematic Risk - The risk inherent to a person or company (think of this as a company that has a competitor that moves in driving down sales). This risk is avoidable through diversification or making bets on many assets preferably uncorrelated.

In this project, we'll be focusing on a specific type of risk called Credit or Default Risk, which has both systemic and unsystemic drivers. The main point is that the drivers of default risk can be measured and analyzed for patterns related to default. As a result, the probability of default for a person or an institution is not random. This is where machine learning can help.

Introduction

Home Credit strives to broaden financial inclusion for the unbanked population by providing a positive and safe borrowing experience.In order to make sure this underserved population has a positive loan experience, Home Credit makes use of a variety of alternative data–including telco and transactional information–to predict their clients' repayment abilities.

For the purposes of this article, we'll focus on the application_train.csv and application_test.csv files, which contain a significant amount of useful information for predicting credit default.This is the main source of information for people that have applied for personal loans including features related to their loan application.

The data is provided by Home Credit, a service dedicated to provided lines of credit (loans) to the unbanked population. Here is the link to the dataset.

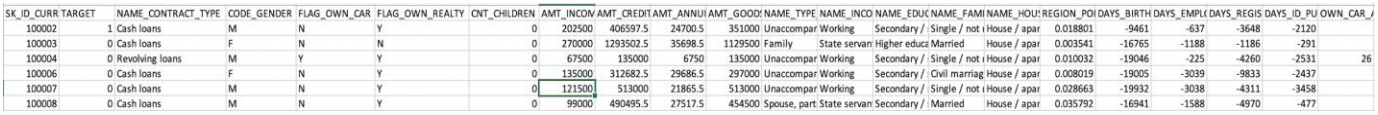
A few points about the data:

- The training data contains 307K observations, which are people applied for and received loans.
- The "TARGET" column identifies whether or not the person defaulted (0 v/s 1).
- The remaining 121 features describe various attributes about the person, loan, and application.
- There are several additional data sets (bureau.csv, previous_application.csv). These auxiliary files contain data that is one-to-many relationship, which would require aggregation (a feature engineering method) and considerably more work for the unknown benefit. Hence, we have chosen to leave analysis of variables included in these additional datasets.

Our analysis delves into how certain factors influence loan default rates. In the end we build a model that predicts if a user is going to default on their loan.

We were more interested in **agile iteration**: Building a good model quickly, then going back to try to improve later. We will be evaluating our model accuracy using ROC AUC curves. We also used a very disciplined Agile process to ensure that we execute critical path tasks in parallel and in small chunks.

Attached below is a representation of the dataset:



SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOM	AMT_CREDIT	AMT_ANNUI	AMT_GOODS	NAME_TYPE	NAME_INCO	NAME_EDUC	NAME_FAMI	NAME_HOUS	REGION_POI	DAYS_BIRTH	DAYS_EMPL(	DAYS_REGIS	DAYS_ID_PU	OWN_CAR_A
100002	1	Cash loans	M	N	Y	0	202500	406597.5	24700.5	351000	Unaccompar	Working	Secondary /	Single / not	House / apar	0.018801	-9461	-637	-3648	-2120	
100003	0	Cash loans	F	N	N	0	270000	1293502.5	35698.5	1129500	Family	State servan	Higher educa	Married	House / apar	0.003541	-16765	-1188	-1186	-291	
100004	0	Revolving loans	M	Y	Y	0	67500	135000	6750	135000	Unaccompar	Working	Secondary /	Single / not	House / apar	0.010032	-19046	-225	-4260	-2531	26
100006	0	Cash loans	F	N	Y	0	135000	312682.5	29686.5	297000	Unaccompar	Working	Secondary /	Civil marriag	House / apar	0.008019	-19005	-3039	-9833	-2437	
100007	0	Cash loans	M	N	Y	0	121500	513000	21865.5	513000	Unaccompar	Working	Secondary /	Single / not	House / apar	0.028663	-19932	-3038	-4311	-3458	
100008	0	Cash loans	M	N	Y	0	99000	490495.5	27517.5	454500	Spouse, part	State servan	Secondary /	Married	House / apar	0.035792	-16941	-1588	-4970	-477	

The target variable is appropriately titled ‘TARGET’ - 1 (defaulters) v/s 0 (non-defaulters)

The input variables include binary variables for gender, car and realty ownership, the number of children, income, annuity, occupation type, marital status etc.

The model is a binary classification supervised learning model.

Descriptive Analysis

In this section, we focus on the distribution of the data to understand how it is structured and what segments of the population the data focuses on.

Here is the distribution of the income variable:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
25650	112500	147150	168798	202500	117000000

The median income level recorded for the data is much higher than we expected (147k USD) when compared to the nation’s median annual household income (50k USD).

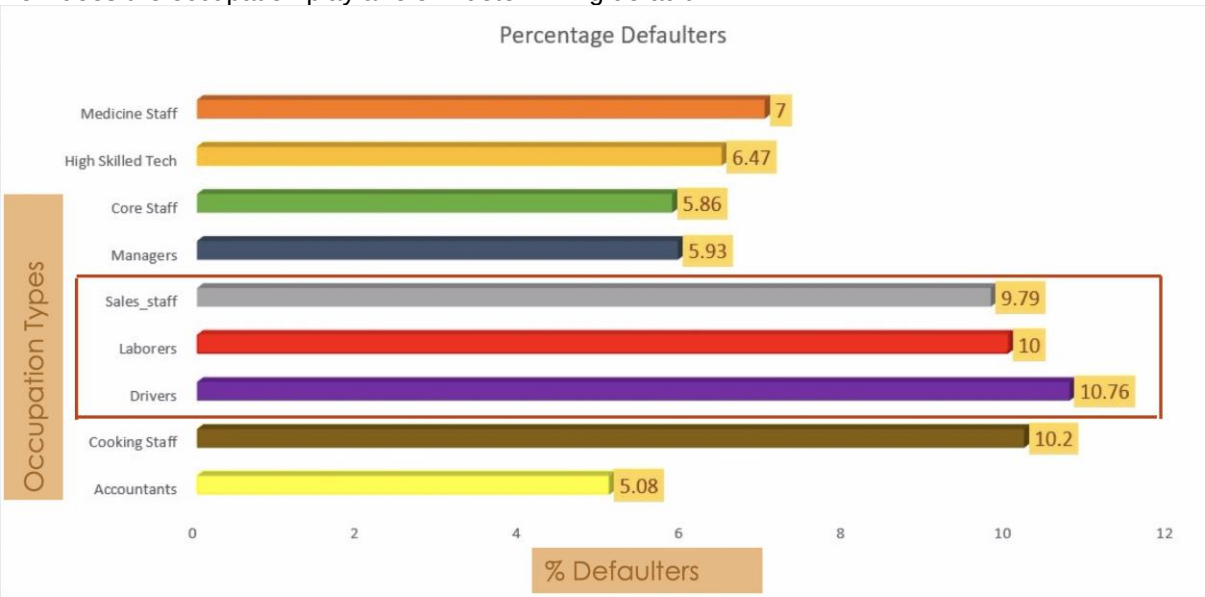
Even though the lowest income recorded is $25,650, the 1st quartile includes a larger amount ($112,500) which indicates the data consists of a large number of high-income earners.

The graph below shows a distribution of annuity amounts that each person has to

Exploratory Data Analysis

We explored different variables included in the dataset to explore relationships and further insights about the data. The following paragraphs explore some of these in more detail:

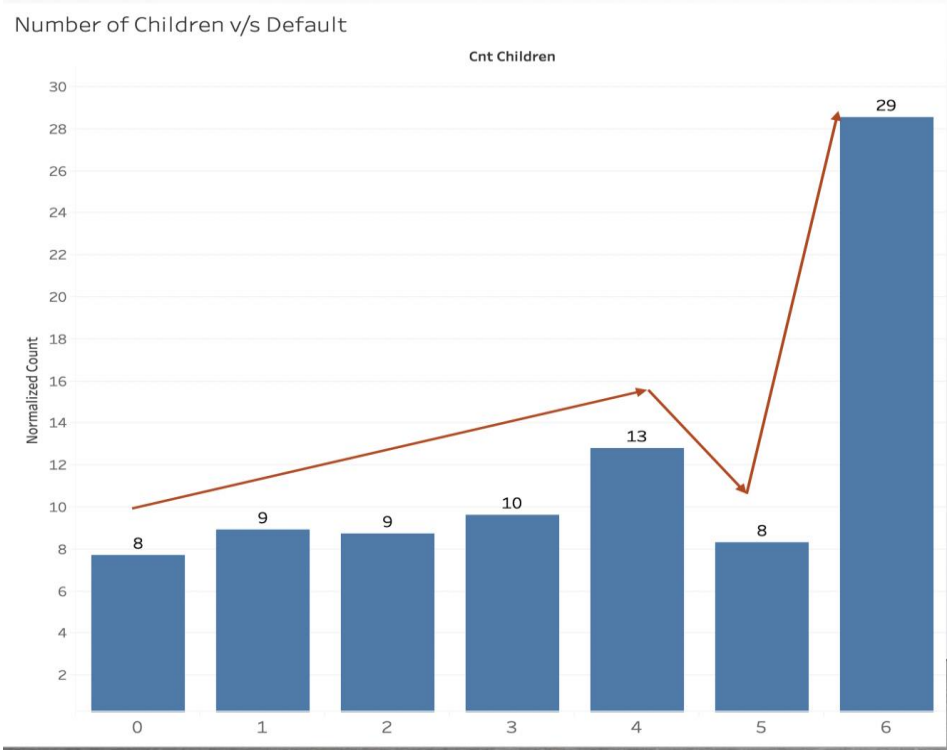
How does the occupation play a role in determining default ?



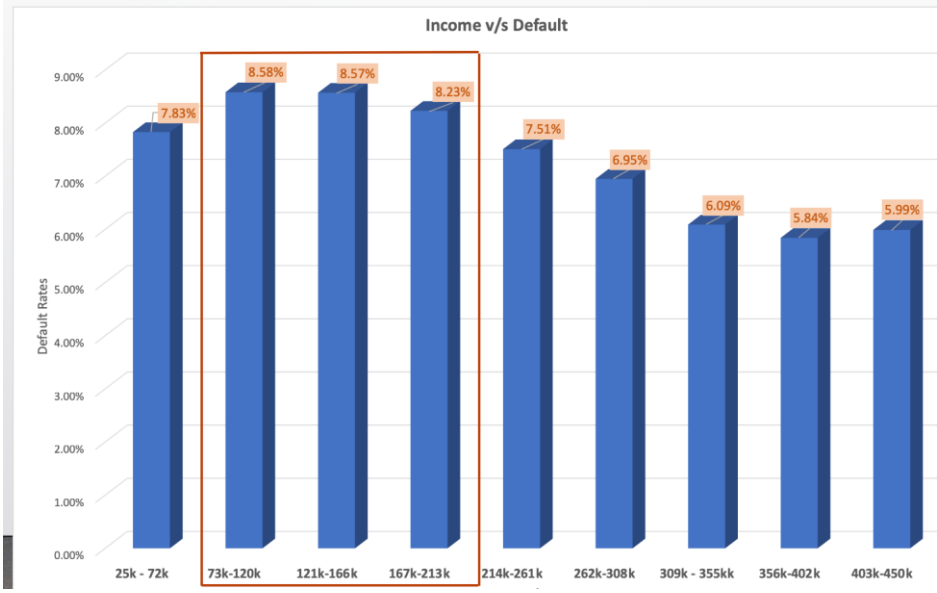
The graph above shows how blue collar employees are more at a risk to default.

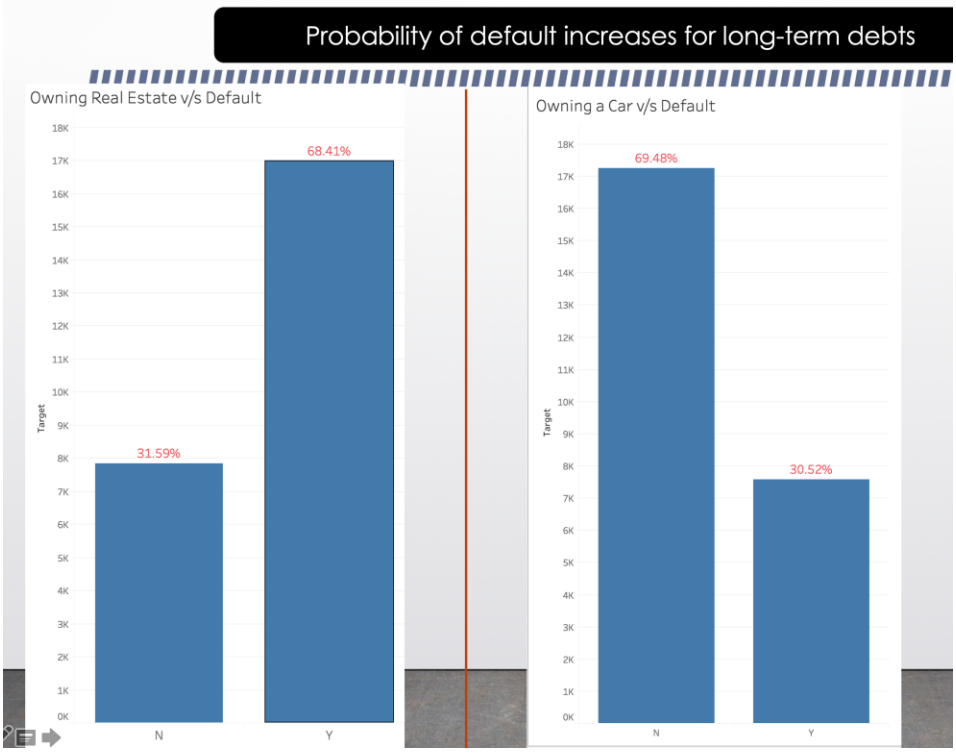
This visualization explores the person's relationship status against their chances of default. Married couples are much more likely to default than all categories of people without dependents combined together.

To further illustrate how having more dependents (not just spouses) puts more financial burden is the comparison of number of kids in a household and the number of cases where they defaulted. Households having more than or equal to 6 kids have a much higher probability of default.



The visualization below explores income levels against the probability of default. The general assumption that having higher income leads to lesser chances of default holds true here. Surprisingly though, the highest rates of default don't happen in the lowest bracket of income but people earning much more than the median household income (50,000 USD for a household of 4 people) are much more likely to default. Additionally, people earning much higher annual incomes don't have an insignificant default rate either.





The visualization above compares default rates with assets that people own.
People owning a car (depreciating asset) are less likely to default than those who don't. This can be partially contributed to the competitive market for auto financing which often finances such purchases for very little interest.

Compare this with the people who took mortgages (appreciating asset) which often have higher interest rates over a longer period of time, and we can see that people owning real estate are more likely to default.

Modify Data

From the Exploratory Data Analysis, we could find that NA plays a major part affecting data quality. Hence, we removed variables which contains NA more than 15%. Also, we have removed variables which are less correlated. We imputed NA values present in important variables, using median. We have converted all the categorical variables to numeric variables, as we have utilized the gradient boosting to predict our defaulters. XGBoost manages only numeric vectors.

To ensure that all data was representative of the segment of population recorded, in all of the visualizations shown above data has been normalized to account for the proportion of people that defaulted from that subsection of the overall population.

Model

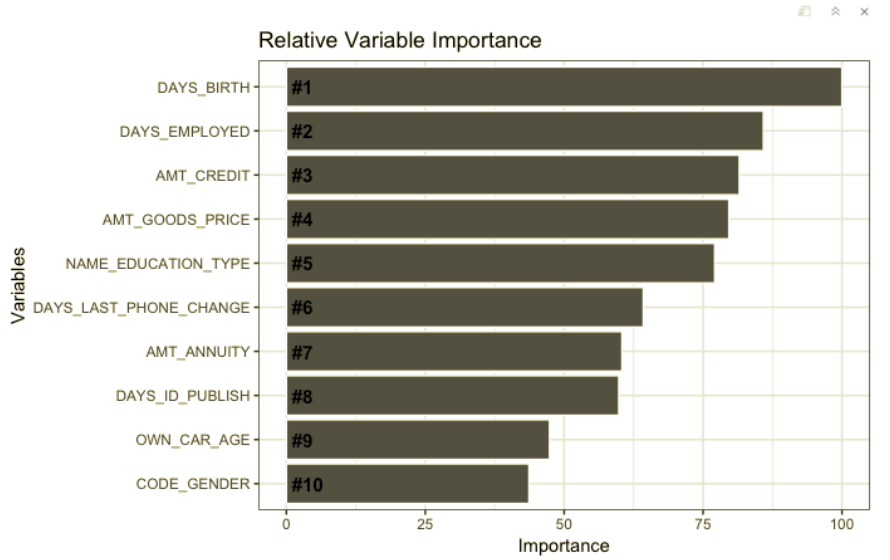
In boosting, the trees are built sequentially such that each subsequent tree aims to reduce the errors of the previous tree. Each tree learns from its predecessors and updates the residual errors. Hence, the tree that grows next in the sequence will learn from an updated version of the residuals.

Formula:
$\mathbf{F_n(x) = F_{n-1}(x) + h_n(x)}$

Why XGBoost?

- Regularization: XGBoost has an option to penalize complex models through both L1 and L2 regularization. Regularization helps in preventing overfitting
- Handling sparse data: Missing values or data processing steps like one-hot encoding make data sparse. XGBoost incorporates a sparsity-aware split finding algorithm to handle different types of sparsity patterns in the data
- Weighted quantile sketch: Most existing tree based algorithms can find the split points when the data points are of equal weights (using quantile sketch algorithm). However, they are not equipped to handle weighted data. XGBoost has a distributed weighted quantile sketch algorithm to effectively handle weighted data

The trained XGBoost model automatically calculates feature importance on our predictive modeling problem.



In the variable importance plot above, we can see that features are named automatically according to their index in the input array. Thus, we can conclude that DAYS_BIRTH has the highest importance relative to all other variables within the training data, whereas CODE_GENDER has the lowest importance.

Evaluation

Plotting the Receiver Operating Characteristic(ROC) curve helped visualize the performance of the binary classifier in predicting the probability of Default Vs No Default. In a ROC curve, the true positive rate (Sensitivity) is plotted as a function of the false positive rate (100-Specificity) for different cut-off points of a parameter.

The area under the ROC curve (AUC) is a measure of how well a parameter can distinguish between two diagnostic groups (experiencing a Default Vs No Default). Based on this plot, the area under the curve for our best model was found to be ~0.89. This shows that a randomly selected observation from the training data set with a '1' (likely to Default) label has a score larger than that for a randomly chosen observation from the group with the '0' (not likely to default) label 89% of the time.

Conclusion

Financial institutions like banks run on lending money to various counter parties where the money they lend is the shareholders money or the depositor’s money which can be pulled out at any instance by the depositor for which the banks need to continually run checks and maintain awareness to lend money in more safer ways as there is no one stop solution in doing so.

Risk in this scenario is transaction with loss. And this transaction with loss comes with probabilities which can be measured and minimized. We utilized the insights which were previously explained to understand the factors which influence a loan borrower to default through analysis and built a predictive model. This model needs to be implemented in the ERP system of the loan sanctioning and approval team to better judge a loan applicant's ability to pay off the loan in the stipulated time.

This implementation is not only to keep a check on the regulatory requirements to manage risk, but also helps the top management to understand the risk exposure not only with respect to credit but also market such that any contingencies or policies which need to be implemented or changed in order to manage the position of the institution in the market. More direct outcomes are the reduction in cost of quality, settlement risks and operating costs resulting in an improvement in the efficiency of the process.

In conclusion not only do all the banks need the understanding of the credit exposure to avoid bankruptcy in the worst-case scenario by having contingencies in place like put hedges in place but also, to avoid the collapse of the banking ecology due to the lack of knowledge on credit debt risk exposure.

Thank You!