Leveraging Advanced Analytics to Enhance Credit Card Underwriting

DS GA 1001 - Final Capstone Project
December 2021

Hriditaa Dekate - hrd259
Aryann Dharamsey - asd484
Kristin Mullaney - kmm9492
Alejandro Sáez - as15796
Joseph Schuman - js12580

NEW YORK UNIVERSITY

Introduction

The purpose of the *Introduction to Data Science* capstone is to apply skills learned throughout the semester to a project with real-world data. In this project, we will utilize hypothesis testing, clustering and classification to unveil patterns within credit card data. First, we will introduce the dataset and explain what we hope to learn from it. Next, we will document our key questions and their corresponding analyses. Finally, we will summarize our conclusions and key findings.

Context and problem statement

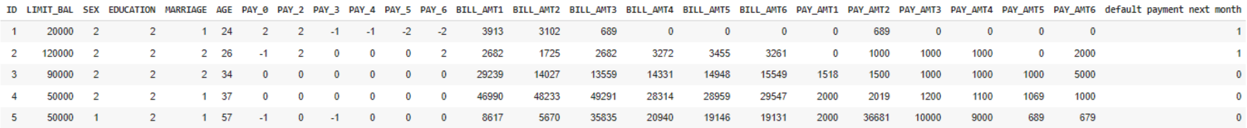
Over the past decade, the field of targeted risk management within the banking industry has been revolutionized by improvements in advanced analytics and the growth of big data. Banks no longer utilize fixed rule-based systems to diagnose whether a client is worthy of a loan, credit line or credit card. Instead, they rely on various models to conduct these tasks. In our project, we will analyze the credit card payments of 30,000 Taiwanese bank clients in service of three objectives:

- **Understanding the credit default behavior** of the client base across demographics with the help of hypothesis testing (e.g. How creditworthy are older clients?)
- **Clustering the client base** to understand the major groups of clients and their relative risk of credit default (e.g. How creditworthy are older, married, college graduates?)
- **Building a machine learning model** that predicts the future likelihood of client default in order to make more informed lending decisions. Specifically, to make decisions on a more granular level as compared to the previous clustering analysis. (e.g. Given everything we know about a particular client's demographic and payment history, will they default on their next payment?)

Data overview

The data used in this study consists of customer level information from a major Taiwanese bank. Its format and contents are summarized below:

- **Source**: UCI ML repository https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients
- **Rows:** 30,000 individual clients
- **Columns:** Credit card limit, sex, education, marital status, age, six month bill history, six month payment history, default in next month flag (1: They did default, 0: They did not)
- **Default rate**: 22.1%
- **Data format:**

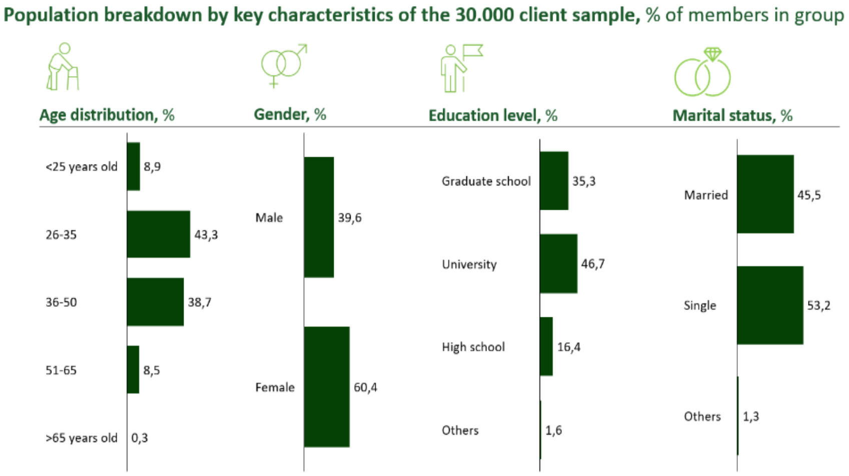


ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6	BILL_AMT1	BILL_AMT2	BILL_AMT3	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6	default payment next month
1	20000	2	2	1	24	2	2	-1	-1	-2	-2	3913	3102	689	0	0	0	0	689	0	0	0	0	1
2	120000	2	2	2	26	-1	2	0	0	0	2	2682	1725	2682	3272	3455	3261	0	1000	1000	1000	0	2000	1
3	90000	2	2	2	34	0	0	0	0	0	0	29239	14027	13559	14331	14948	15549	1518	1500	1000	1000	1000	5000	0
4	50000	2	2	1	37	0	0	0	0	0	0	46990	48233	49291	28314	28959	29547	2000	2019	1200	1100	1069	1000	0
5	50000	1	2	1	57	-1	0	-1	0	0	0	8617	5670	35835	20940	19146	19131	2000	36681	10000	9000	689	679	0

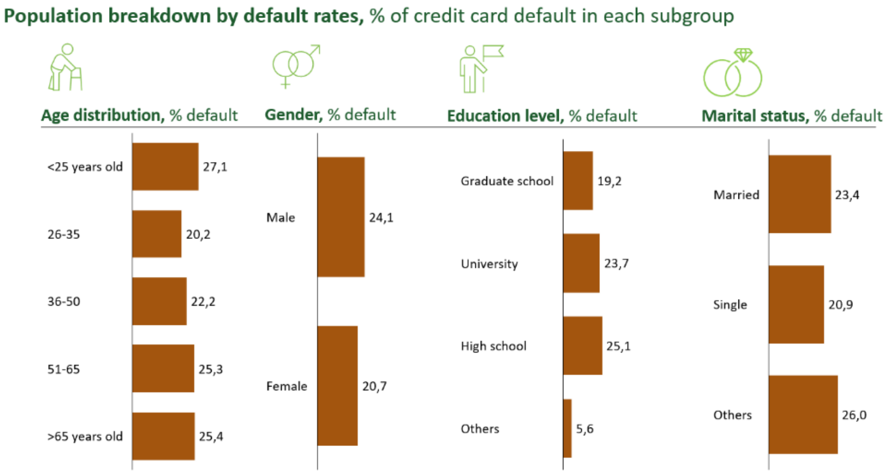
Question 1: Are some client demographics more likely to default than others?

We first approached this problem by performing exploratory data analysis to see how different demographics were represented within our data set.

We found that of our 30,000 clients, over 80% were between the ages of 26 and 50. We also found that about 60% were female, and that over 80% had a university degree or higher. The split between married and single clients was about even.



6,636 clients defaulted on their next credit card payment, accounting for about 22.1% of total customers. The following diagram depicts the default rates for various demographic subgroups.



From our exploratory analysis, we found that younger clients have the highest default rates, that men default more than women, and that rates of default decrease with higher levels of education.

Our next step was to employ a null hypothesis testing framework to see if these differences could reasonably be accounted for by randomness alone. We first assumed that the distributions for defaulting were identical across all demographic and other categorical subgroups. We then chose the appropriate statistical test to see how likely we were to sample our particular data under this assumption. Our results are listed in the table below.

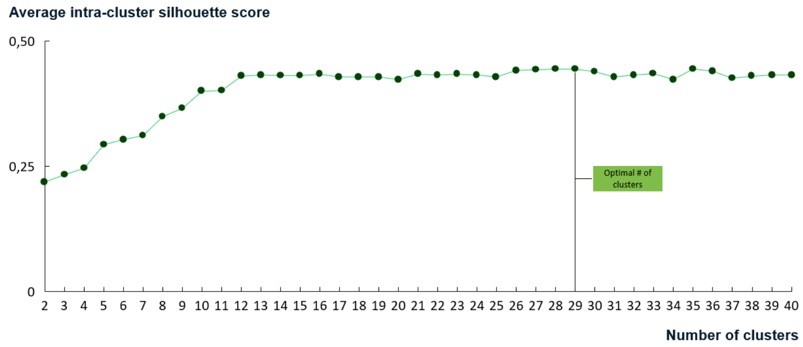
Question	*Approach*	*Result*
Do default rates differ for clients across education levels?	Executed a Chi-square test to assess whether the observed differences in default rates across education levels are likely due to chance.	P value: 1.86e-17 **It is reasonable to conclude that the observed differences in default rates across education levels are unlikely due to chance alone.**
Do default rates differ for clients across age groups?	Executed a Chi-square test to assess whether the observed differences in default rates across age groups are likely due to chance.	P value: 4.66e-13 **It is reasonable to conclude that the observed differences in default rates across age groups are unlikely due to chance alone.**
Do default rates differ for clients with higher vs lower credit limits?	Executed a Chi-square test to assess whether the default rates differ for clients with higher credit limits vs those with lower credit limits.	P value: 7.62e-117 **It is reasonable to conclude that clients with higher credit limits default at a different rate than clients with lower credit limits.**

Question 2: Are there clusters of clients with higher propensities to default?

To answer this question, we performed a KMeans clustering algorithm on all of our demographic indicators to find 29 differentiated groups of customers. After descriptively characterizing each cluster, we assessed the statistical significance of the observed default rates across all of the clusters.

We scaled our customer demographic data to have zero mean and a standard deviation of one to better highlight intrinsic client groups based on key factors, and more accurately analyze their respective default rates. In order to choose the most appropriate number of clusters we have performed a silhouette analysis as depicted in the right-hand side.

Silhouette analysis to determine optimal cluster number, avg intra-cluster silhouette score



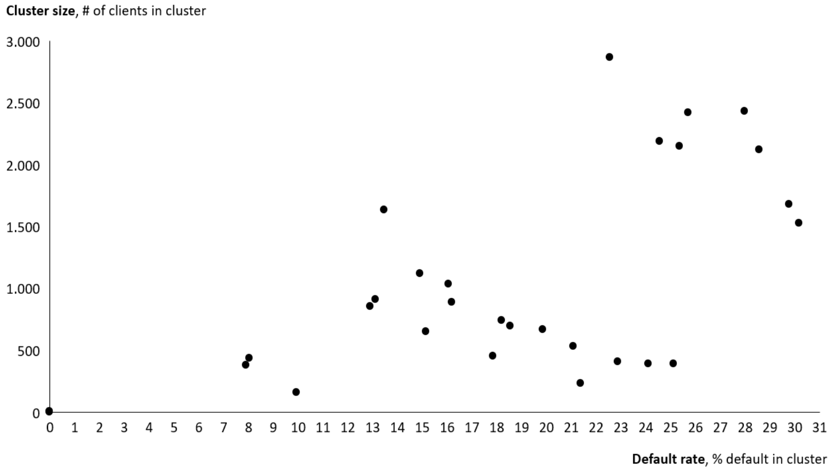
Seeing how the optimal number of clusters by this metric is 29, we display the characteristics of the 5 clusters with the most number of customers on the right hand side. We displayed the demographics with which we clustered as well as other characteristics.

Descriptive analysis of characteristics of largest clusters

Cluster number	# Clients in cluster	Average age	% Graduate school	Average credit card limit	% default rate in cluster
1	2.865	30,6	71	85.123	23
2	2.433	36,7	0	78.337	28
3	2.418	29,5	0	70.616	26
4	2.187	40,5	36	101.152	25
5	2.152	32,2	62	74.196	25

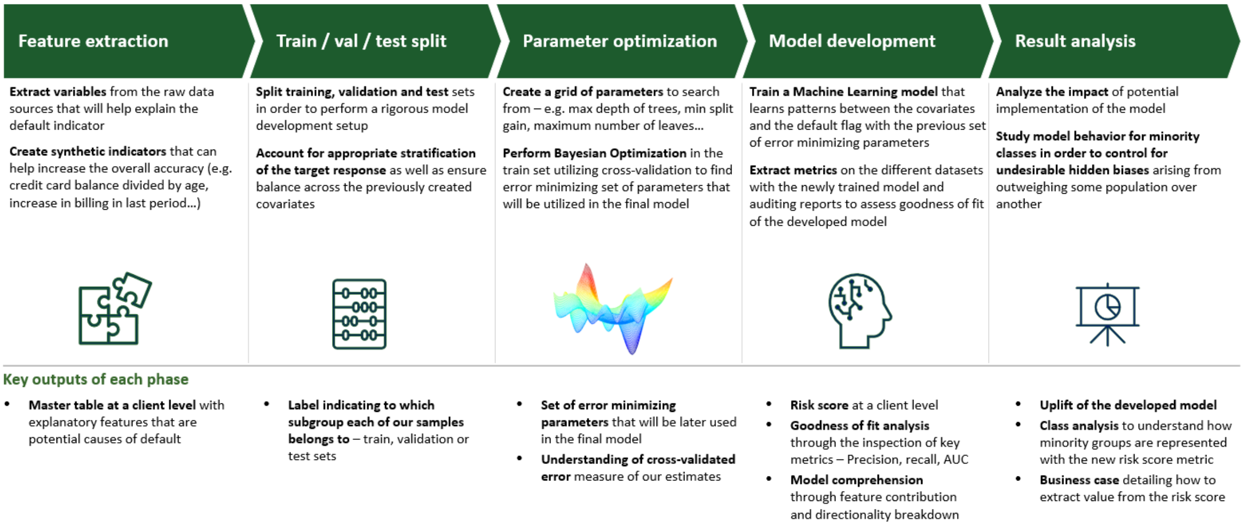
A natural immediate question is whether this descriptive analysis can be used to influence lending decisions. In the graph to the right, we observe the discrepancy in terms of default rate across all 29 clusters. We observe that there are clusters with 0% default rates while others have default rates of 31%. This procedure to identify risky clients could be used to influence lending decisions. In the next chapter we will explore how to calculate a risk metric at an individual level and not only at a group level.

Discrepancy of default rate across clusters



Question 3: How accurately can we predict individual-level probabilities of default?

Up until now, we have analyzed descriptive relations in the data and found subgroups of clients with higher propensities to default. However, iIn order to be able to influence underwriting decisions, i.e. who gets a credit card and who doesn't, we need a risk metric which quantifies an individual's likelihood of default. To be thorough, we built three different models and compared how well each could predict when a client would default. For each model, we employed a 5-step approach to creating our risk-score as depicted below.



Feature extraction

Extract variables from the raw data sources that will help explain the default indicator

Create synthetic indicators that can help increase the overall accuracy (e.g. credit card balance divided by age, increase in billing in last period…)

Train / val / test split

Split training, validation and test sets in order to perform a rigorous model development setup

Account for appropriate stratification of the target response as well as ensure balance across the previously created covariates

Parameter optimization

Create a grid of parameters to search from – e.g. max depth of trees, min split gain, maximum number of leaves…

Perform Bayesian Optimization in the train set utilizing cross-validation to find error minimizing set of parameters that will be utilized in the final model

Model development

Train a Machine Learning model that learns patterns between the covariates and the default flag with the previous set of error minimizing parameters

Extract metrics on the different datasets with the newly trained model and auditing reports to assess goodness of fit of the developed model

Result analysis

Analyze the impact of potential implementation of the model

Study model behavior for minority classes in order to control for undesirable hidden biases arising from outweighing some population over another

Key outputs of each phase

- **Master table at a client level** with explanatory features that are potential causes of default

- **Label indicating to which subgroup each of our samples belongs to** – train, validation or test sets

- **Set of error minimizing parameters** that will be later used in the final model
- **Understanding of cross-validated error** measure of our estimates

- **Risk score** at a client level
- **Goodness of fit analysis** through the inspection of key metrics – Precision, recall, AUC
- **Model comprehension** through feature contribution and directionality breakdown

- **Uplift of the developed model**
- **Class analysis** to understand how minority groups are represented with the new risk score metric
- **Business case** detailing how to extract value from the risk score

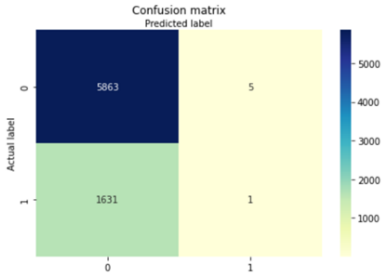
Logistic Regression

The first attempt at modelling credit risk goes through using a logistic regression. The main advantage of this method is the ability to audit it (no black box) and explain the factors this model relies upon. However, and as we will see, its main pitfall is lack of predictive power when benchmarked with more refined models

Feature Extraction: The goal of feature extraction is to isolate any features that might impact an individual's risk of default. Since we could not be sure that any of the demographic or payment history features were completely independent of default risk, we incorporated all of these features in our model.

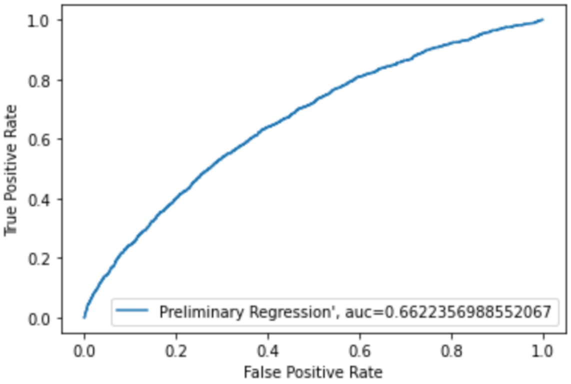
Train/Val/Test split: We put 25% of our data into a testing set and used the remaining 75% to train our model.

Parameter Optimization: We used the logistic regression function within Scikit-Learn to find the regression coefficients that would maximize the predictive ability of the model.

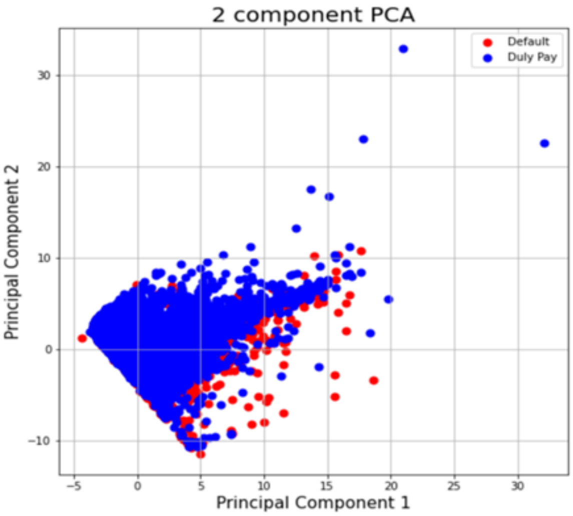


Model Development: When applying our model to our test data, we achieved an accuracy of 0.78, a precision of 0.16 and a recall of 0.006. The breakdown of our results can be seen in the confusion matrix to the right.

To account for the fact that our results rely on the threshold that our model uses to determine what should be classified as a 1 (will default) or 0 (won't default), we studied the AUC-ROC curve. The AUC was 0.66, indicating that our model does have some predictive ability.



Result Analysis: We standardized our data and performed a PCA to investigate why our model is struggling with precision and recall. We found that with two principal components, there is a lot of overlap in component characteristics. This could be confusing our model. We reimplemented the logistic regression, this time with standardized data and the help of the PCA. This revised model has an accuracy of 0.78, a precision of 0.5, and a recall of 0.0006. Our precision improved, but our recall remained very low.



Extreme Gradient Boosting decision trees

Having studied the performance of a logistic regression on our dataset, we will move on to try extreme gradient boosting. As discussed earlier, the main advantage of this model is the increased level of predictive power at the cost of loss of interpretability. However, gradient boosting machines have been established as a main mechanism for such tasks as credit underwriting for their capacity to capture non-monotonic relations between the input features and the target response. As earlier, we will go through the 5 steps followed.

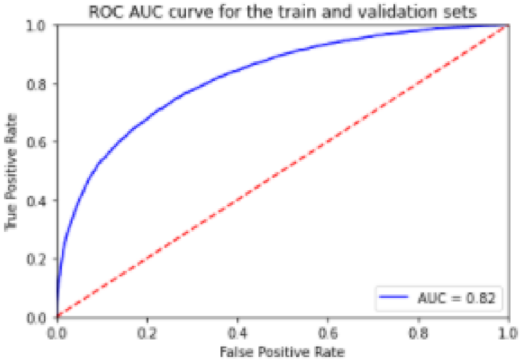
Feature Extraction: For this model, we again are going to use all of the client demographic and credit history data that is available to us. We one-hot encoded synthetic variables that represent each possible marital status and education level. For example, by separating each education level into its own column containing only 0's and 1's, the model can more accurately account for the non-ordinal nature of education.

Train/Val/Test split: We used Bayesian optimization to find the optimal set of parameters for our gradient boosting machine using a 5-fold cross validation. Once the optimal set of parameters was found we validated with 10% of our data and trained with the remaining 90% to have at our disposal the final model.

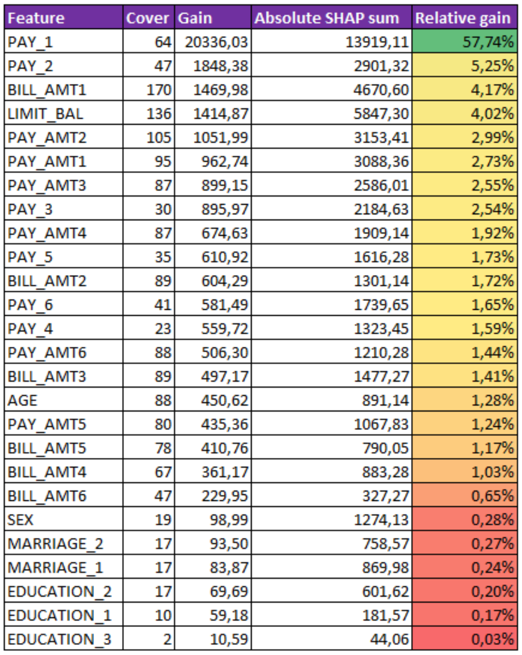
Parameter Optimization: Since gradient boosting has a broader set of parameters to tune for, we have executed a bayesian optimization search with 5-fold cross validation using a well established framework: https://github.com/fmfn/BayesianOptimization. Essentially, a range is specified for each and every parameter within a grid and a number of iterations are performed to find the error-minimizing set of parameters. In our specific case, the optimal set of parameters found are the ones itemized below:

Parameter	*Value*
Bagging fraction	`0.971358121572873`
Feature fraction	`0.6567649610573583`
Lambda L1	`2.7651285211727434`
Lambda L2	`2.805707392349148`
Max depth	`11`
Min child weight	`12.992542951810567`
Min split gain	`0.054149713453980786`
Number of leaves	`23`

Model Development: Having found an error-minimizing set of parameters, we move on to the model development itself. We have trained a gradient boosting machine with lightgbm's implementation (https://github.com/microsoft/LightGBM) with a holdout set of 10% of our data to determine early stopping, The out of sample results of our model improve the earlier found classifier with the logistic regression significantly as can be observed below:



Result Analysis: Having improved our previous classifier, the natural question is: on which factors does the gradient boosting machine rely on the most. While in the case of the logistic regression this was a straightforward question to answer, here we will use "out-of-model" methods such as gain calculation and SHAP values to quantify the relative contributions of different features to the outcomes of our model. The table below illustrates the relative contribution of different factors according to various metrics:



Feature	Cover	Gain	Absolute SHAP sum	Relative gain
PAY_1	64	20336,03	13919,11	57,74%
PAY_2	47	1848,38	2901,32	5,25%
BILL_AMT1	170	1469,98	4670,60	4,17%
LIMIT_BAL	136	1414,87	5847,30	4,02%
PAY_AMT2	105	1051,99	3153,41	2,99%
PAY_AMT1	95	962,74	3088,36	2,73%
PAY_AMT3	87	899,15	2586,01	2,55%
PAY_3	30	895,97	2184,63	2,54%
PAY_AMT4	87	674,63	1909,14	1,92%
PAY_5	35	610,92	1616,28	1,73%
BILL_AMT2	89	604,29	1301,14	1,72%
PAY_6	41	581,49	1739,65	1,65%
PAY_4	23	559,72	1323,45	1,59%
PAY_AMT6	88	506,30	1210,28	1,44%
BILL_AMT3	89	497,17	1477,27	1,41%
AGE	88	450,62	891,14	1,28%
PAY_AMT5	80	435,36	1067,83	1,24%
BILL_AMT5	78	410,76	790,05	1,17%
BILL_AMT4	67	361,17	883,28	1,03%
BILL_AMT6	47	229,95	327,27	0,65%
SEX	19	98,99	1274,13	0,28%
MARRIAGE_2	17	93,50	758,57	0,27%
MARRIAGE_1	17	83,87	869,98	0,24%
EDUCATION_2	17	69,69	601,62	0,20%
EDUCATION_1	10	59,18	181,57	0,17%
EDUCATION_3	2	10,59	44,06	0,03%

As one can note, all 26 variables contribute to the model results but with varying levels of importance. The payment amount in the last month and penultimate month seem to be the ones contributing the most to the model result as one might expect.

Key Findings

As a summary of the previous analyses, in this chapter we intend to itemize what we have learned from the study as well as highlight some potential steps forward that could help enhance the work presented here.

- ❖ **Differences in creditworthiness across education levels**: We have descriptively seen that education is a determining factor in the creditworthiness of a client as shown by the Chi-Square test that we ran
- ❖ **Differences in creditworthiness across age buckets**: We also saw through the use of a Chi-Square test that the different creditworthiness of the different age ranges that we observed were unlikely due to chance alone, thus allowing us to reject the null hypothesis.
- ❖ **Differences in credit balance across riskiness profile**: Through the use of a Chi-square test we saw that clients with different levels of credit limit present a likely not random disparity in default rates
- ❖ **Client clustering**: Through a KMeans clustering we identified well-separated groups of clients with a significant discrepancy in default rates
- ❖ **Individual level risk modeling**: Fine tuning the previous descriptive analyses, we have developed two classifiers to predict probability of default at a client level with a resulting AUC of 82 points.

Some potential next steps that could enhance our work are itemized below:

- **Identify which regulations credit card underwriters are subject to**: Since the banking industry is subject to strict regulations, some of the indicators used might not be appropriate from a regulatory standpoint - gender, marital status… A potential enhancement would be to double check which attributes can and cannot be used in order to have a ready-to-deploy model.

- **Include additional sources of information**: In order to have a more holistic view of the client additional sources of information could be included such as income, assets under management or lifetime value within the bank.

- **Develop a business case where the optimal threshold-cutoff is calibrated**: In this study we have created classifiers that output the probability of default. However, underwriters need a hard metric which tells them whether a credit card should be granted or not. Incorporating the cost of default information we could get a calibration for what the optimal cutoff point is.