Beedareddy Jeevan Sai Reddy – M12368208

Aniket Sunil Mahapure – M12910618

Feature Selection through Hypothesis Testing

## Abstract

Hypothesis testing is widely used in Marketing analytics to assess impact of campaigns. Doing the same would provide insights on strategies to develop. In this project, we would like to use hypothesis testing for feature selection i.e. selecting best features before building the model. We will focus on testing difference of medians of selected features (only continuous variables) amongst classification levels. Parametric bootstrap will be used for features on which assumptions about distribution can be made and non-parametric bootstrap for others. Wald test will be used to do hypothesis testing

## Introduction

Classification problems usually involve hundreds of features and requires feature engineering to select best few amongst them. This helps in improving computational efficiency and in improvements to model. Usually, in industry algorithms like PCA are used for feature engineering. In this project, we propose to demonstrate usage of hypothesis testing for the same. We will be using credit cards default dataset in this project

## Problem Statement

One of the banks in Taiwan has around 30000 credit card customers and wants to predict customers who would be defaulting in the upcoming month. To do the same, we are provided with customers’ demographic data, billing history data and payment history data.

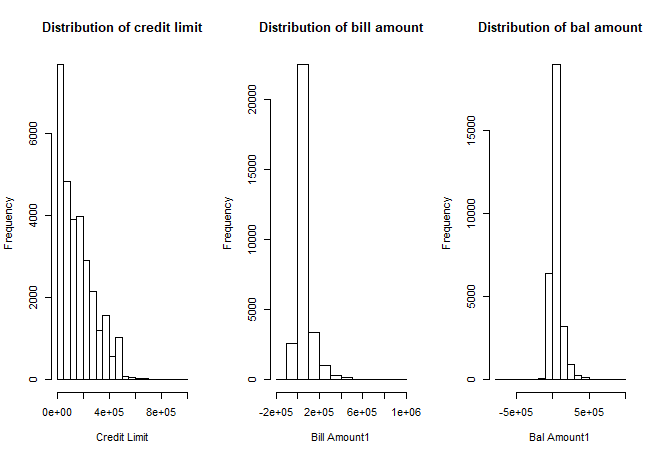
Below is the table explaining available variables. By plotting distributions, we decided upon bootstrap methodology (parametric / non-parametric bootstrap)

|  |  |  |
| --- | --- | --- |
| ***Variable*** | ***Description*** | ***Methodology*** |
| LIMIT\_BAL | Credit card usage limit | Non Parametric |
| zero\_duly | 1 if zero duly payments are done else 0 | Parametric |
| Atleast 1 duly | 1 if at least 1 duly payments are done else 0 | Parametric |
| Atleast 2 duly | 1 if at least 2 duly payments are done else 0 | Parametric |
| Atleast 3 duly | 1 if at least 3 duly payments are done else 0 | Parametric |
| Atleast 4 duly | 1 if at least 4 duly payments are done else 0 | Parametric |
| Atleast 5 duly | 1 if at least 5 duly payments are done else 0 | Parametric |
| Atleast 6 duly | 1 if at least 6 duly payments are done else 0 | Parametric |
| BILL\_AMT1 | Billed amount for past month | Non Parametric |
| BILL\_AMT2 | Billed amount for second last month | Non Parametric |
| BILL\_AMT3 | Billed amount for third last month | Non Parametric |
| BILL\_AMT4 | Billed amount for fourth last month | Non Parametric |
| BILL\_AMT5 | Billed amount for fifth last month | Non Parametric |
| BILL\_AMT6 | Billed amount for sixth last month | Non Parametric |
| PAY\_AMT1 | Amount paid for past month | Non Parametric |
| PAY\_AMT2 | Amount paid for second last month | Non Parametric |
| PAY\_AMT3 | Amount paid for third last month | Non Parametric |
| PAY\_AMT4 | Amount paid for fourth last month | Non Parametric |
| PAY\_AMT5 | Amount paid for fifth last month | Non Parametric |
| PAY\_AMT6 | Amount paid for sixth last month | Non Parametric |
| default payment next month | Default Flag in next month 1 / 0 | Non Parametric |
| BAL\_AMT1 | Balance bill for past month (Billed - Paid) | Non Parametric |
| BAL\_AMT2 | Balance bill for second last month (Billed - Paid) | Non Parametric |
| BAL\_AMT3 | Balance bill for third last month (Billed - Paid) | Non Parametric |
| BAL\_AMT4 | Balance bill for fourth last month (Billed - Paid) | Non Parametric |
| BAL\_AMT5 | Balance bill for fifth last month (Billed - Paid) | Non Parametric |
| BAL\_AMT6 | Balance bill for sixth last month (Billed - Paid) | Non Parametric |

## Non-Parametric Bootstrap

Upon observing distributions of billed amount, found that most them are right skewed. Couldn’t assume any specific distribution and hence went ahead with non-parametric bootstrap

Below is the distribution of three sample variables:



Below are Null and Alternate hypotheses that would be tested for. Will be using Wald test for the same

H0 – Null Hypothesis – Median of (eg: Bill amount, credit limit etc) for defaulted customers = Median of (eg: Bill amount, credit limit etc) for non-defaulted customers

H1 – Alternate Hypothesis – Median of (eg: Bill amount, credit limit etc) for defaulted customers != Median of (eg: Bill amount, credit limit etc) for non-defaulted customers

1. ***Defining theta function:***

Since we are testing for Median, Theta function would also be median

1. ***Generating Bootstrap:***

We will be generating two bootstraps and find bootstrap median for both samples (Defaulted and Non – Defaulted customers)

After generating the same, we will be finding out difference (Defaulted – Non Defaulted) between the generated bootstrap medians. This would be used to find standard error of difference of medians

1. ***Generating point estimate:***

Point estimates of medians will be generated for both samples and difference would be calculated from the same

1. ***Generating confidence intervals:***

Using point estimate of difference of medians and standard error calculated, we will be finding out confidence interval of the same

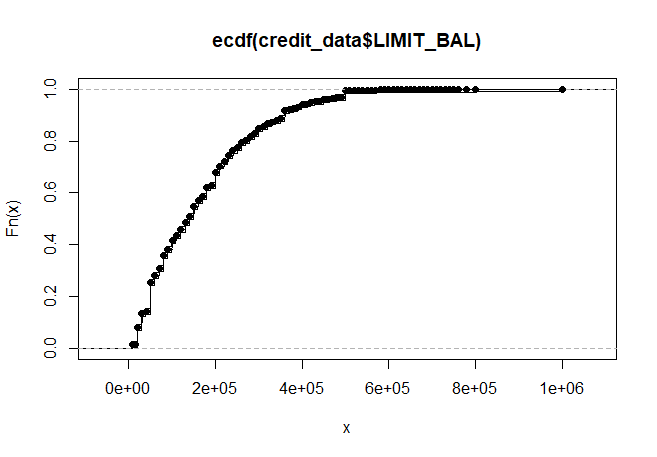
The same process would be done for all selected variables under non-parametric bootstrap to obtain confidence intervals of differences of medians. If confidence interval contains zero, then can reject Null hypothesis, else can’t reject the same

Below is the list of variables on which we were able to reject Null Hypothesis point estimates of median difference and confidence intervals

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***Variable*** | ***median\_hat\_diff*** | ***se\_median\_diff*** | ***lower\_limit*** | ***upper\_limit*** |
| LIMIT\_BAL | -60000 | 2091.883356 | -64183.76671 | -55816.23329 |
| BILL\_AMT1 | -2934.5 | 714.2089858 | -4362.917972 | -1506.082028 |
| BILL\_AMT2 | -1360 | 635.2886441 | -2630.577288 | -89.42271188 |
| BILL\_AMT6 | 1349.5 | 410.089303 | 529.321394 | 2169.678606 |
| PAY\_AMT1 | -823.5 | 57.12208635 | -937.7441727 | -709.2558273 |
| PAY\_AMT2 | -714 | 53.0736956 | -820.1473912 | -607.8526088 |
| PAY\_AMT3 | -778 | 32.90429925 | -843.8085985 | -712.1914015 |
| PAY\_AMT4 | -734 | 33.37536995 | -800.7507399 | -667.2492601 |
| PAY\_AMT5 | -765 | 37.81106479 | -840.6221296 | -689.3778704 |
| PAY\_AMT6 | -706 | 32.95900037 | -771.9180007 | -640.0819993 |
| BAL\_AMT3 | 1210 | 407.9736229 | 394.0527542 | 2025.947246 |
| BAL\_AMT4 | 1306 | 407.3983573 | 491.2032855 | 2120.796715 |
| BAL\_AMT5 | 2347 | 469.467314 | 1408.065372 | 3285.934628 |
| BAL\_AMT6 | 3482.5 | 579.2114322 | 2324.077136 | 4640.922864 |

## Ecdf for credit limit balance

From above table found that median difference of credit limit balance to be little high. Hence, to understand distribution of the same, we will be plotting ecdf for credit limit balance



Could see that there is one data point whose value is very high and few data points whose value is less. Anyway, we are not going to treat the same given the possibility of occurrence of such instances.

Could also see that confidence band is very narrow indicating better confidence intervals

## Callouts

1. Could see that median difference of credit limit is around -60000. This indicates that customers with low credit limit might default more
2. Difference of payments across all months is negative irrespective of Billed amount. This indicates that tendency to default has started much prior to default date

## Parametric Bootstrap

In statistics, bootstrapping is any test or metric that relies on random sampling with replacement. Bootstrapping allows assigning measures of accuracy (defined in terms of bias, variance, confidence intervals, prediction error or some other such measure) to sample estimates. This technique allows estimation of the sampling distribution of almost any statistic using random sampling methods

Following columns are created using past monthly payment records (PAY\_0, PAY\_2……PAY\_6).

|  |  |
| --- | --- |
| ***Column Name*** | ***Description*** |
| no of duly payments | No of duly payments done |
| zero\_duly | 1 if zero duly payments are done else 0 |
| Atleast 1 duly | 1 if at least 1 duly payments are done else 0 |
| Atleast 2 duly | 1 if at least 2 duly payments are done else 0 |
| Atleast 3 duly | 1 if at least 3 duly payments are done else 0 |
| Atleast 4 duly | 1 if at least 4 duly payments are done else 0 |
| Atleast 5 duly | 1 if at least 5 duly payments are done else 0 |
| Atleast 6 duly | 1 if at least 6 duly payments are done else 0 |

We will be testing if ratio of no. of default customers to total customers is equal across newly created binary variables.

For column ‘zero duly’

P1 = ratio of no. of default customers to total customers where zero duly = 1

P2 = ratio of no. of default customers to total customers where zero duly = 0

Ho: P1 – P2 = 0

H1: P1 – P2 ≠ 0

Since these columns have Bernoulli distribution, We have developed 3000 samples using sample function in R for P1 and P2 and their respective MLE values of proportions. Then confidence intervals were calculated for difference of P1 and P2. If 0 does not lie in respective 95% confidence interval, then we can reject Ho.

Similar process is repeated for each column for hypothesis testing to decide its important for model creation.

Summary of hypothesis testing for all columns

Result = 1 if 0 lies in 95% confidence interval else 0

|  |  |  |  |
| --- | --- | --- | --- |
| **Column Name** | **CI** | **P1 - P2** | **Result** |
| Atleast 1 duly | 0.07139308, 0.08450033 | -0.07794671 | 0 |
| Atleast 2 duly | -0.08466735, -0.07122607 | -0.07519343 | 0 |
| Atleast 3 duly | -0.08186618, -0.06852069 | -0.07533246 | 0 |
| Atleast 4 duly | -0.08190039, -0.06876453 | -0.0772917 | 0 |
| Atleast 5 duly | -0.08377942, -0.07080397 | -0.07803975 | 0 |
| Atleast 6 duly | -0.08430428, -0.07177523 | -0.08422206 | 0 |

## Callouts

1. By looking at results, we can say that we can reject the Ho for all of the above columns since zero doesn’t lie in 95% confidence interval.
2. Since magnitude of difference is almost equal for all the columns, we can say there is a pattern in customer’s past behavior of payments. If a customer did not make a duly payment in past months, he is more likely to not pay duly in future months as well and vice versa.
3. There is a significant difference between P1 and P2 for all columns.
4. All these columns should be used for building prediction model.

## Bayesian Approach

Till now, we have calculated confidence intervals using frequentist approach. Same can be done using Bayesian approach as well.

In order to do the same, we would need prior and likelihood function. Let’s assume non-informative prior i.e. Beta(1,1) and calculate likelihood function for Bernoulli. Using the same let’s calculate posterior for p and find confidence intervals of above variables

Below is the comparison of Confidence intervals calculated using Parametric bootstrap and Bayesian approach

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Parametric Bootstrap CI** | **Bayesian CI** |
| Atleast 1 duly | -0.07139308, -0.08450033 | -0.07139126, -0.08441629 |
| Atleast 2 duly | -0.08466735, -0.07122607 | -0.06878033, -0.08155518 |
| Atleast 3 duly | -0.08186618, -0.06852069 | -0.06890609, -0.08170633 |
| Atleast 4 duly | -0.08190039, -0.06876453 | -0.07099364, -0.08364718 |
| Atleast 5 duly | -0.08377942, -0.07080397 | -0.07190225, -0.08426146 |
| Atleast 6 duly | -0.08430428, -0.07177523 | -0.07803840, -0.09065353 |

* As we can see, confidence intervals from both methods are very similar for every column.
* Frequentist and Bayesian approach are giving us similar results.
* Non-informative prior gave same results when sufficient sample size is present.

## Conclusions

1. This is new approach to perform feature selection and we leveraged non-parametric bootstrap and parametric bootstrap for the same
2. Were able to reduce number of required variables to model from 27 to 20 (Eliminated 5 through non-parametric bootstrap and 0 through parametric bootstrap)
3. Doing the same in real time scenario would help in improving computational efficiency many folds

## Bibliography

<https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients> – Source of dataset under study

All of statistics: A Concise Course in statistical inference

Lecture Notes – BANA 7031 – Probability Models – MS Business Analytics, University of Cincinnati