Consolidated Report

Predict Credit card Default

Lakpa Sherpa

Project Proposal

- **What is the problem you want to solve?**

  A Bank's main service is to provide loans, and credit cards to people who are in need of money to either invest on a business, purchase assets, pay school tuitions or other reasons. It is a great source of income for a bank until the loan or credit is set as a default. Then, the bank has to spend time and money to take legal actions against the debtor. The money may or may not be collected depending on the circumstances. However, Banks can take precautionary steps to pin out applicants who are at risk to default.

- **Who is your client and why do they care about this problem? In other words, what will your client do or decide based on your analysis that they wouldn't have done otherwise?**

  This project is useful for bank's risk management officers to find out high risk level credit card applicants and to recommend new loans and credit card to low risk loan/credit applicants who are a potential client for the bank. This project will predict if a client will default a credit card loan, so the bank can take precautionary measures to minimize the possibility. The bank can find out the variables which affects the client's behaviors or possibility to default. With the model we create from this project and the necessary client's data and informations feed into our model, we can predict if the client will or will not default. This will help the bank to make decision to issue credit card.

- What data are you using? How will you acquire the data?

  I am going to use a data set of Taiwanese credit card holders which was donated to UCI Machine Learning Repository for study and research purposes. The data set consists of

30,000 rows of records with 25 columns. You can download the excel file to acquire the data from the website.

The Data Set includes following data from the client which will help us to create different machine learning models:-

- Client ID: Unique number used as identity of the clients.
- X1: Amount of the given credit (NT dollar): Limit_Balance
- it includes both the individual consumer credit and his/her family (supplementary) credit.
- X2: Gender (1 = male; 2 = female).
- X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
- X4: Marital status (1 = married; 2 = single; 3 = others).
- X5: Age (year).
- X6 - X11: History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows:
- X6 = the repayment status in September, 2005
- X7 = the repayment status in August, 2005
- X8 = the repayment status in July, 2005
- X9 = the repayment status in June, 2005
- X10 = the repayment status in May, 2005
- X11 = the repayment status in April, 2005.
- The measurement scale for the repayment status is:-
- -1 = pay duly
- 1 = payment delay for one month
- 2 = payment delay for two months
- 3 = payment delay for three months
- 4 = payment delay for four months
- 5 = payment delay for five months
- 6 = payment delay for six months
- 7 = payment delay for seven months
- 8 = payment delay for eight months
- 9 = payment delay for nine months and above.
- X12-X17: Amount of bill statement (NT dollar).
- X12 = amount of bill statement in September, 2005

- X13 = amount of bill statement in August, 2005
- X14 = amount of bill statement in July, 2005
- X15 = amount of bill statement in June, 2005
- X16 = amount of bill statement in May, 2005
- X17 = amount of bill statement in April, 2005
- X18-X23: Amount of previous payment (NT dollar).
- X18 = amount paid in September, 2005
- X19 = amount paid in August, 2005
- X20 = amount paid in July, 2005
- X21 = amount paid in June, 2005
- X22 = amount paid in May, 2005
- X23 = amount paid in April, 2005
- Y(default Payment) 1 for yes and 0 for no

- **Briefly outline how you'll solve this problem. Your approach may change later, but this is a good first step to get you thinking about a method and solution.**

- **Data Wrangling:**
- Import the file to the database as a pandas DataFrame
- Clean the data using different manipulating techniques to tidy and rearrange the data
- Change column names.
- Make separate columns for different gender, education level, and marital status.

- **Data Visualization:**
- Use Seaborn to visualize the data with different kinds of plot
- Plot age vs repayment status to see any pattern
- Plot marital status vs repayment status for different months
- Plot gender vs repayment status for different months
- Plot education vs repayment status for different months

- Model:
- Logistic Regression
- Decision Tree

- ➢ Random Forest
- ➢ Extra Trees
- ➢ Gradient Boosting
- ➢ Support Vector Machine
- ➢ Neural Networks(Optional)

- What are your deliverables? Typically, this includes code, a paper, or a slide deck.

The deliverables for this project would be a complete presentation of the project with all the documentations, visualizations and python codes on notebook. All the model created to predict default on credit card loan. The final submission will include the introduction, the steps and methods involved in this project from the beginning to the end of the project. Finally, present how we evaluated the accuracy of different models created, and recommend a model.

Data Wrangling

First, I imported required packages that we would need for Data Wrangling. Then, I used pandas read_excel module to import excel file directly from the website of UCI Machine Learning Repository. The excel file had two column labels for each column, so we skipped the first row to ignore the label that we do not need. Then, I converted that imported OrderedDict to a pandas DataFrame. The column label/names were inconsistent, so i changed the column labels with better column labels. Columns for specific month included month name, and space in between from column default payment next month was changed to default_payment. Then, I checked to see if there were any missing value and outliers.

First Approach (update and modify bad data and Outliers) :
In column ‘education’, the value 1 represent graduate school, 2 represent university, 3 represent high school, and 4 represent others. However, there were values 5, 6 and 0 which would be a mistake or those value does not represent any level of education. I located those values and replaced those value with 4 and kept them under the category 4 representing others. Here, I assumed that the best possible value for these bad data would be others. Others meaning any other education level except 1 graduate, 2 university, and 3 high school. This is the reason why i choose others would be the best possible value.

Addition to this, I will deal with the data by using other techniques like using mode and median to replace the bad data and outliers, and drop/delete data to exclude such bad data from affecting our study. The first and important approach would be to consult or ask with other team members if they know exactly why we have such bad data.

In column 'marital_status' the value 1 represent married, 2 represent single, and 3 represent others. However, there was value 0 present in the column which did not represent any category. That value may be a missing value or a mistake or others. Here, I located those value and replaced those value with 3 which represent others assuming other to be the best possible value.

There are six column for repayment status for month september, august, july, june, may, and april. In each column, value -1 represent pay duly, 1 for payment delay one month, 2 for payment delay two months,........8 for delay 8 months and 9 for delay 9 month or above. There are value 0 and -2, which are bad data. I checked to see what those value were in previous months. Then, located those bad data to change them to -1, which represent duly paid.

- What kind of cleaning steps did you perform?
  Checked for missing data
  Checked for inconsistent column names
  Checked for Outliers
  Checked column data types

- How did you deal with missing values, if any?
  There are no missing values for this dataset.

- Were there outliers, and how did you handle them?
  There were outliers in the age column, the balance amount columns for each month, and the amount paid columns for each month. We checked for outliers using Box Plot Diagram and 1.5*IQR rule. Both of them proved that outliers were present. I created a copy of current dataframe to update and modify the outlier values from those columns. Data higher or greater than Q3+1.5*IQR is replaced by Q3+1.5*IQR, and data lower than Q1-1.5*IQR is replaced by Q3-1.5*IQR. I did not change the outliers from the age column.

  Second Approach (Remove/Drop bad data and outliers):
  In the second approach, I dropped all the rows with bad data and outliers instead of estimating the best possible value to replace them. First, I would locate any bad data and outliers, then I drop them to exclude them from the analysis.

  Third Approach (Replacing bad data with mode value and Outliers with median value) :
  In the third approach, I located any bad data and replaced them with their mode value. Mode value is the most repeated and majority of the value. I did not use mean and

median, because mode is a better choice for categorical data. Then I replaced any outlier values to their median value. Median value is the middle value or mid-point of the data set.

Project: Capstone Project 1
Data Story

Approach 1

After we wrangled and cleaned the dataset, we started to explore the data in detail. The first step was to see the count and distributions of different variables from the dataset.

- How many cardholders are defaulters?
  We found that the 77.88% of the cardholders(23,364) did not default and 22.12% of the cardholders (6,636) default.
- How are the cardholders divided by gender?
  We have 60% female and 40% male clients.
- What are the education level, and Which education level does the most of the cardholders belong to?
  Most of our cardholder have University level education for their highest level of education. We have 35% with Graduate level education, 46% with University level, 16% with High School level, and 1.5% with Others as level of education.
- How many cardholders are married and how many are single?
  We have 53% married, 45% single, and rest as others.
- What age group is the majority of the cardholders?
  Most of our cardholders are of age group 20 to 40. This exploration can provide us with the demographic of different variables.

The main focus of this project is to create different Machine Learning Models to predict default, so let's find out some insights using different data visualization. We drew barplots to compare different variables with the default cardholders. From each plot, we learned which sub-variable effects default. We found that University level cardholders default more than other education level. From marital_status with default plot, we found that both married and single have very close number of default, and others have a very low default. Age group with default shows that age group 20 to 30 default more than any other age-group. We found that some cardholders have negative balance. This means that some cardholders are paying the bank more than their balance or some transaction of purchase may have been refunded to the credit card. When we compared balance column with paid column for the month of september, we saw that there is a positive linear relationship between the balance and paid columns except for some balances. Looking at the plot, those balances may be paid by automatic payment every month.

From the analysis and visualizing the dataset, we found that the maximum number of defaulter are female, with University level education, and age between 20 and 30. When we compared male and female for each age group with default, we saw that more female of age group 20-30 default more, and both older male and older female after age group 50-80 default less. When we group gender and education column, we see that 2,2 or Female with University level education has the maximum default. When we compared gender, education and marital status with default, we found that a female, university level, and married cardholder has the maximum default.

Approach 2

After we wrangled and cleaned the dataset, we started to explore the data in detail. The first step was to see the count and distributions of different variables from the dataset.

- How many cardholders are defaulters?
  We found that about 72% of the cardholders(10,315) did not default and about 28% of the cardholders (4,042) default.
- How are the cardholders divided by gender?
  We have 60% female and 40% male clients.
- What are the education level, and Which education level does the most of the cardholders belong to?
  Most of our cardholder have University level education for their highest level of education. We have 30% with Graduate level education, 49% with University level, 19% with High School level, and 0.33% with Others as level of education.
- How many cardholders are married and how many are single?
  We have 53% married, 45% single, and rest as others.
- What age group is the majority of the cardholders?
  Most of our cardholders are of age group 20 to 40. This exploration can provide us with the demographic of different variables. Most of our clients are between age 20 and 30.

The main focus of this project is to create different Machine Learning Models to predict default, so let's find out some insights using different data visualization. We drew barplots to compare different variables with the default cardholders. From each plot, we learned which features or variable effects default. When we compared different features with default, we found that the gender, age, education, and marital status has effect on cardholders default. We found that University level cardholders default more than other education level. From marital_status with default, we found that both married and single have very close number of default, and others have a very low default. Age group with default shows that age group 20 to 30 default more than any other age-group. We found that some cardholders have negative balance. This means that some cardholders are paying the bank more than their balance or some transaction of purchase may have been refunded to the credit card. When we compared balance column with paid column for the month of september, we saw that there is a positive linear relationship between the balance and paid columns except for some balances. Looking at the plot, those balances may be paid by automatic payment every month.

From the analysis and visualizing the dataset, we found that the maximum number of defaulter are female, and with University level education. When we compared male and female for each age group with default, we saw that female of age group 20-30 are the most default. When we group gender and education column, we see that 2,2 or Female with University level education default more than other group clients. When we compared gender, education and marital status with default, we found that group female, university level, and married cardholder had maximum default.

Approach 3

After we wrangled and cleaned the dataset, we started to explore the data in detail. The first step was to see the count and distributions of different variables from the dataset.

- How many cardholders are defaulters?
  We found that the 77.88% of the cardholders(23,364) did not default and 22.12% of the cardholders (6,636) default.
- How are the cardholders divided by gender?
  We have 60% female and 40% male clients.

- What are the education level, and Which education level does the most of the cardholders belong to?
  Most of our cardholder have University level education for their highest level of education. We have 35% with Graduate level education, 48% with University level, 16% with High School level, and 0.4% with Others as level of education.
- How many cardholders are married and how many are single?
  We have 53% married, 45% single, and rest as others.
- What age group is the majority of the cardholders?
  Most of our cardholders are of age group 20 to 40. This exploration can provide us with the demographic of different variables.

The main focus of this project is to create different Machine Learning Models to predict default, so let's find out some insights using different data visualization.

We drew barplots to compare different variables with the default rate of total cardholders. From each plot, we learned which sub-variable has greater default. When we compared gender with default, we found that the female cardholders default more than male. We found that University level cardholders default more than other education level. From marital_status with default plot, we found that both married and single have very close rate of default, and others have a very low default rate. Age group with default shows that age group 20 to 30 default more than any other age-group. We found that some cardholders have negative balance. This means that some cardholders are paying the bank more than their balance or some transaction of purchase may have been refunded to the credit card. When we compared balance column with paid column for the month of september, we saw that there is a positive linear relationship between the balance and paid columns except for some balances. Looking at the plot, those balances may be paid by automatic payment every month.

From the analysis and visualizing the dataset, we found that the maximum number of defaulter are female, with University level education, and age 20-30. When we compared male and female for each age group with default, we saw that more female of age group 20-30 default more than male, and both older male and older female after age group 50-80 default less. When we group gender and education column, we see that 2,2 or Female with University level education default more than other group clients. When we compared gender, education and marital status with default, we found that group female, university level, and married cardholder had maximum default.

Exploratory Data Analysis

From data visualization and exploration, I found that the gender, education and age affected the default_payment. The collected data shows that more female default than male card holders. The plot shows that the card holders with university level education have the maximum default rate, and the plot show some correlation between these variables. Age group with default rate plot shows that the younger the age group, higher the default rate. It shows a inverse relationship between age-group and default rate.

According to the data, I found that the female card holders default more than male card holders. To check if this did not happen by chance, I performed a population proportion difference hypothesis T-test. The null hypothesis is that the population male default mean and population female default mean is equal. The alternative hypothesis is population male mean and population female default mean is not equal. We got a t-value of 6.85, and a very low p-value. The test suggests that the male mean default and female mean default is not equal, because the p-value calculated is less than the level of significance 0.05.

I also conducted a Chi-squared test to evaluate the relationship between gender and default payment, education-level and default payment, marital_status and default payment, and age-group and default payment. First, I evaluated the relationship between gender and default payment. The null hypothesis is that gender and default_payment are independent of each other or not related. The alternative hypothesis is that gender and default_payment are not independent or related. I used the level of significance alpha 5%. The p-value calculated from the chi-squared test was very low less than 0.001. Because the p-value is less than 0.05, we reject the null hypothesis and suggest the alternative hypothesis. Therefore the gender and default payment are related and dependent to each other.

Then, I performed the chi-squared test for education and default_payment. The p-value for this test was very low, less than 0.001. The result is significant at $p<0.05$, so we will reject the null hypothesis. We have enough evidence to support the alternative hypothesis that the education level and default_payment are related and dependent on each other.

Then, I performed the chi-squared test for marital_status and default_payment. The null hypothesis for this test is that marital_status and default_payment are not related. The alternative hypothesis for this test is that marital_status and default_payment are related. After the calculation, p-value we got is approximately 0.0000007, which is less than 0.001. The level of significance is 0.05. Because the p-value is less than 0.05, we reject the null hypothesis that the marital_status and default_payment are not related. We suggest the alternative hypothesis that the marital_status and default_payment are related.

Finally, I compared age-group and default_payment with chi-squared test. The null hypothesis is that age-group and default_payment are not related. The alternative hypothesis is that age-group and default_payment are related. The test suggested that the age-group and default_payment are dependent and related to each other. The p-value is approximately 0.0000003, which is less than the level of significance 0.05. The evidence suggest the alternative hypothesis and rejects the null hypothesis.

Machine Learning In Depth Analysis

In this project, I am going to compare different classification models with each other to build a best performing model to predict the credit card default. Firstly, I am going to handle class imbalance present in the dataset using different resampling methods. Then, I will find the best method of handling class imbalance for the dataset we have. Then, I will use that method to handle the class imbalance. The new data set will be used to train different models.

First, I created dummies features for all the categorical data from the dataset. All the categorical data must be converted into machine readable language which is numerical. Then, our classification model will be able to understand and learn from that dummy feature. We have 3 categorical features i.e. gender, education, and marital_status. When we convert the categorical features to dummy variable, the features would be subdivided into its category and given 1 or 0 values for yes and no. The status column for different months had values less than 0 like -1,-2,-3 and those numbers were changed to 0. -1 represented that the client paid in time.

-2 represented that the client paid 1 months early. -3 represented that the client paid 2 months in advance.

Then, I imported all the necessary packages that i would be using for the machine learning analysis. Then, I divided the dataset into input ‘X’ and label ‘y’. By the help or using input ‘X’, we will find out, predict, or classify the label ‘y’. Then I scaled the input features of X using robust scaler because or dataset contains some outlier values in paid and balance column. Then, I divided the X and y into training and testing data set. The training set is 70% of our total data and training set is 30% of the data. Our dataset has class imbalance which needs to be taken cared to build a better model which would not over/underfit when learning for classification. We need to handle class imbalance before we start building different model. I used different methods of handling class imbalance techniques to find the best method for our data set. I tried balancing the class weight, undersampling and oversampling, and ensemble method. Balancing the class weight will add weight to the lower class by copying the data from the lower class to balance the weight of the both class. Undersampling will ignore or not include some data from the upper class in our dataset to balance with the lower class. Oversampling will copy the data from the lower class and add them to our dataset to balance with upper class. Ensemble method will combine two techniques together to handle the class imbalance.

I build a logistic regression model as our base classification model. Then we used different methods of handling class imbalance techniques on our dataset. I calculated different scoring metrics to measure the performance of our model. I stored the score of different class imbalance handling methods to compare which method was the best. I used F1 score to measure the performance of our model because we have class imbalance in our dataset. Higher the f1 score, better would be our model.

Approach 1

The ensemble method SMOTEEEN which is a combination of SMOTE and Edited Nearest Neighbour method had the best f1 score. According to the F1 Score of the classification model, the SMOTE+ENN method to handle the class imbalance performed the best. So we will use SMOTE+ENN to handle class imbalance, then compare different models with each other.

Now, I am using SMOTEEEN as the resampling method and building different models with the new resampled dataset. I build Logistic Regression, Decision Tree, GaussianNB, Random Forest, and Support Vector Machine classifier. Logistic Regression has the best f1 score compared to all other classification model, so I will choose Logistic Regression for this project. Since the F1-score is significantly better than a random classifier, I would recommend this Logistic Regression model to predict credit card default. Let's try to tune all our model to see if we can improve the performance and then decide which classification model would be the best.

Hyperparameter is a user provided argument or input which effects the performance of the model. Hyperparameter tuning is done by providing a set of hyperparameter values for the machine learning model, and the model will choose its best parameter which achieves the best performance score. I used GridSearchCV and RandomizedSearchCV to perform the Hyperparameter Tuning. I used 5 fold cross validation to find the best f1 score for the all the model. After tuning the hyperparameter of all different model, we can see that the f1 score for Decision Tree had decreased after hyperparameter tuning, and the f1 score for Random Forest has improved and increased. Logistic Regression, GaussianNB, and SVM has not changed even after tuning. Logistic Regression has the best f1 score compared to all other classification model, so I will choose Logistic Regression for this project. Since the F1-score is significantly better than a random classifier, I would recommend this Logistic Regression model to predict credit card default.

Approach 2

For approach 2, the Edited Nearest Neighbour method to handle the class imbalance performed the best according to the F1 score, so I used ENN to handle class imbalance, then compare different models with each other. Random Forest has the best f1 score compared to all other classification model, so I will choose Random Forest for this project. Since the F1-score is significantly better than a random classifier, I would recommend this Random Forest model to predict credit card default.

Approach 3

According to the F1 Score of the classification model, the Edited Nearest Neighbour method to handle the class imbalance performed the best, so I used ENN to handle class imbalance for approach 3, then compared different models with each other. Random Forest has the best f1 score compared to all other classification model, so I will choose Random Forest for this project. Since the F1-score is significantly better than a random classifier, I would recommend this Random Forest model to predict credit card default.

The highest f1 score is 56% for approach 2 Random Forest Classifier, followed by 54% for approach 1 Logistic Regression Classifier, and 53% for approach 3 Random Forest Classifier.