|  |  |
| --- | --- |
|  |  |
| CREDIT ONE |  |
|  |  |
|  | PREPARE AND EXPLORE DATA -PYTHON |
|  | SUGITHA DEVARAJAN |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  | OVERALL PROBLEMIncrease in customer default rates - This is bad for Credit One since we approve the customers for loans in the first place.Revenue and customer loss for clients and, eventually, loss of clients for Credit One | |  |
|  | **Investigative Questions:**   1. How do you ensure that customers can/will pay their loans? Can we do this?   SOME LESSONS LEARNT:   1. We cannot control customer spending habits 2. We cannot always go from what we find in our analysis to the underlying "why" 3. We must on the problem(s) we can solve: What attributes in the data can we deem to be statistically significant to the problem at hand? 4. What concrete information can we derive from the data we have? 5. What proven methods can we use to uncover more information and why? | Image |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  | IDENTIFY WHICH CUSTOMER ATTRIBUTES RELATE SIGNIFICANTLY TO DEFAULT RATE | **STEPS TAKEN TO ANALYZE AND PERFORM DATA CLEANUP:**   1. Imported required libraries – pandas, numpy, matplotlib, pandas\_profiling etc. 2. Imported the data from SQL 3. Initial analysis showed (data.head()) that the data need clean up 4. Removed Header (invalid row) and the second row was the actual header. 5. Found 202 rows of duplicate – removed duplicates 6. Dropped ID column since that will cause unnecessary interference in visualization and when pivoting/grouping. 7. With data.info found the data was not all integer, some where object. Decided to use **label encoder** to make all column values integer. |  |
|  | DATA EXPLORATION TO IDENTIFY SIGNIFICANT ATTRIBUTES TO DEFAULT RATE **METHODS USED:**   * PANDAS PROFILING * SEABORN CHARTS – PAIRPLOT, CATPLOT, FACET GRID ETC * BOX PLOTS * HISTOGRAM * GROUPING * FILTER * PIVOT * DISCRETIZATION | |  |

### Data.info after cleanup

Table

Description automatically generated

### Attribute info after label encoding:

Graphical user interface, application, Teams

Description automatically generated

## PANDAS PROFILING:

Generates profile reports from a panda Data Frame. The pandas df.describe() function is great but a little basic for serious exploratory data analysis. pandas\_profiling extends the pandas Data Frame with df. profile\_report () for quick data analysis.

For each column the following statistics - if relevant for the column type - are presented in an interactive HTML report:

* **Type inference**: detect the types of columns in a Data frame.
* **Essentials**: type, unique values, missing values
* **Quantile statistics** like minimum value, Q1, median, Q3, maximum, range, interquartile range
* **Descriptive statistics** like mean, mode, standard deviation, sum, median absolute deviation, coefficient of variation, kurtosis, skewness
* **Most frequent values**
* **Histograms**
* **Correlations** highlighting of highly correlated variables, Spearman, Pearson and Kendall matrices
* **Missing values** matrix, count, heatmap and dendrogram of missing values
* **Duplicate rows** Lists the most occurring duplicate rows
* **Text analyses** learn about categories (Uppercase, Space), scripts (Latin, Cyrillic) and blocks (ASCII) of text data

### Some important observations of credit one data from pandas profiling:

* There are no missing values or duplicates
* Variables types are numerical, Boolean (Sex and default), and categorical (education and marriage).
* Limit balance – min is 10,000 and max is 1,000,000
* Past history gives some valid data with high percentage of zeros. Zeros means use of revolving credit.
* Bill amount seems to have high correlations

Graphical user interface, table

Description automatically generated

Graphical user interface, application

Description automatically generated

Graphical user interface

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

Chart, timeline

Description automatically generated

### Other Observations from EDA:

* 1. We have less data on default

Chart, bar chart

Description automatically generated with medium confidence

* 1. Found some anomalies in Limit\_Bal using box plot after $600K

Chart, box and whisker chart

Description automatically generated

* 1. Bill amount has negative value.

Chart

Description automatically generated

* 1. Pay\_X seems to be important with values greater than 0 which can help predict default.

Chart, box and whisker chart

Description automatically generated

* 1. Facet Grid - which is one of the most interesting functions in the Seaborn library!

It allows you to visualize data sets with lots of columns especially categorical columns

* Which sex defaults more?

Not defaulted are more by females than males

Shape

Description automatically generated with medium confidence

* Limit\_Bal

Anything over 750K to 1M is not defaulted.

All default is in the lower limit balance which is less than 250K

Shape

Description automatically generated with medium confidence

* PAY\_X

Not default is seen more when the Pay\_X is greater than 0 meaning who missed payments.

Month over month it has slowly increased if you see from April to September

A picture containing text, indoor, window, screen

Description automatically generated

* AGE

Age 30 to 40 are high in not defaulting

Shape

Description automatically generated with medium confidence

* PAY\_AMTX

Pay\_AMT zeros are in more % than other values and tends to default.

A picture containing icon

Description automatically generated

* 1. With catplot we can confidently say that when customers are well educated then they don’t default.

(0 is graduate School, 3 is University, 1 is High School and 2 is others)

A picture containing bar chart

Description automatically generated

* 1. Age Discretization – When age is divided into bins I see the default and not default behave very similar.

Chart, bar chart

Description automatically generated

* 1. Limit\_Bal have more defaults in the lower values.

Chart, box and whisker chart

Description automatically generated

* 1. Limit\_Bal discretization – Again this says the same thing as box plot above where the default is more in the lower limit balance.

Chart, bar chart

Description automatically generated

### Report Focus:

### Did you learn anything of potential business value from this analysis?

Yes, I found that even though correlations are week among variables there are some with high correlation like the Bill\_Amt which need to be used in my next step of data modeling to predict. I also found Limit\_BAL and PAY\_X are other variables which will help me predict in my data modeling.

### What are the main lessons you've learned from this experience?

I learnt about pandas profiling and Facet grid along with pd.melt. I also explored about correlation interpretation where I found about the significance test. I need to understand more about the hypothesis test.

<https://towardsdatascience.com/eveything-you-need-to-know-about-interpreting-correlations-2c485841c0b8>

### What recommendations would you give based on your findings?

To predict the default customers, we need to model using the following feature variables

LIMIT\_BAL, PAY-X, BILL\_AMTX, and PAY\_AMTX