

ANL252

# End-of-Course Assessment - July Semester 2022

**Python for Data Analytics**

|  |  |
| --- | --- |
| Course | ANL252 |
| TMA Title | ECA |
| Name | VAN LAL REM |
| SUSS PI No | Z2172902 |
| Date | 4 Sep 2022 |

**Question 1:**

The credit facility dataset provided can be classified into two variables. Categorical and numerical variables.

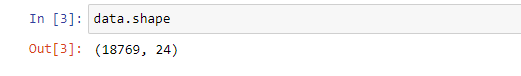
The categorical variables in this dataset are Gender, Education, Marital and Rating. These are the data that are descriptive in nature and can be divided into categories.

The numerical variables in this dataset are Limit, Balance, Income, Age, S(n), B(n), R(n). This variable consists of numbers and can be measured or quantified.

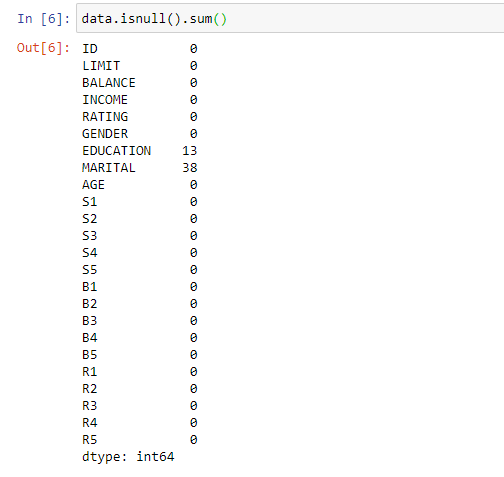
# Question 2:

Data pre-processing tasks:

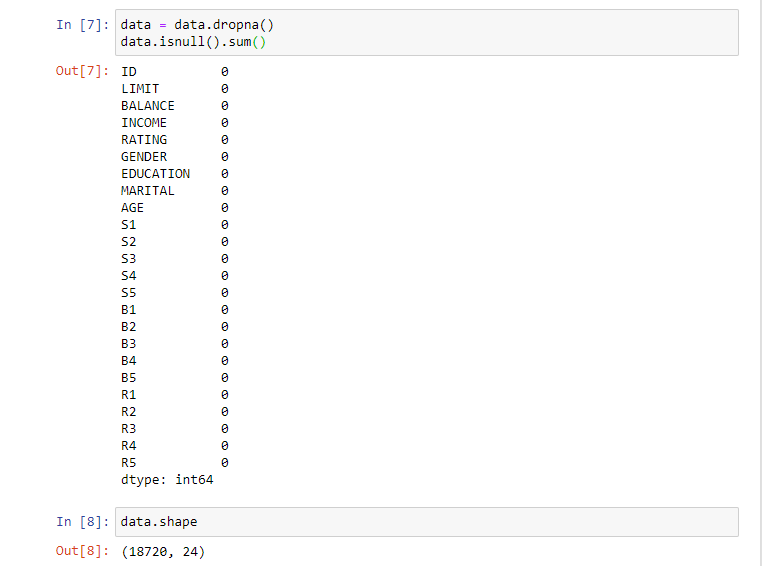
* **Step-1: Missing Value Treatment.**
* There are 18769 rows and 24 columns in the dataset.



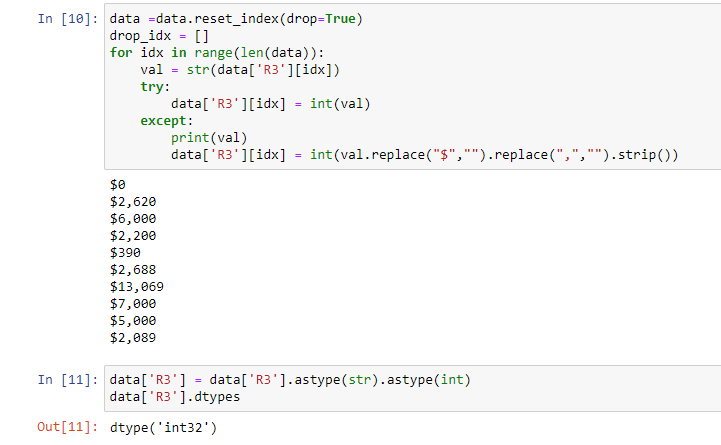
* After checking missing values for each column, there are 13 rows for EDUCATION and 38 rows for MARITAL.



* Because missing rows are less among total rows of the dataset, we drop those rows by using pandas.dataframe.dropna() function.
* As a result, we get 18720 rows that do have not any missing values for each column.

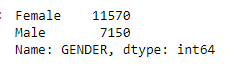


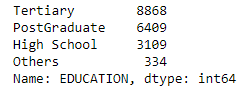
* **Step-2: Pre-Processing for R3 column.**
* There are some string type values in the R3 column and they should be cleaned and converted into numerical type so that the R3 column type can be an integer.
* As there are “$’ in some rows, it can be removed and cleaned.

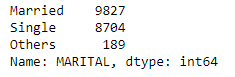


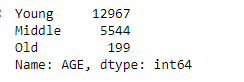
* **Step-3: Convert Numerical Values into Categorical Values for some Columns.**
* As there are some categorical values but they are representing integer values, it should be converted into meaningful names and then it can be encoded again.
* For example, in the GENDER column, 0 is Male, and 1 is Female, it can be replaced by using map() function.

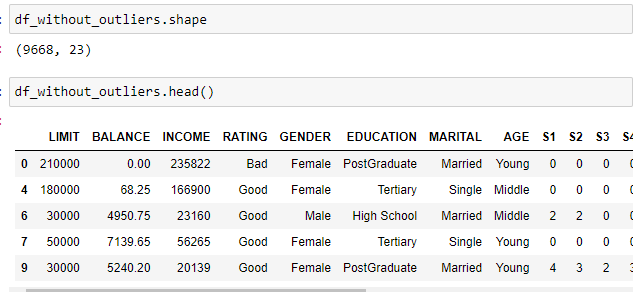


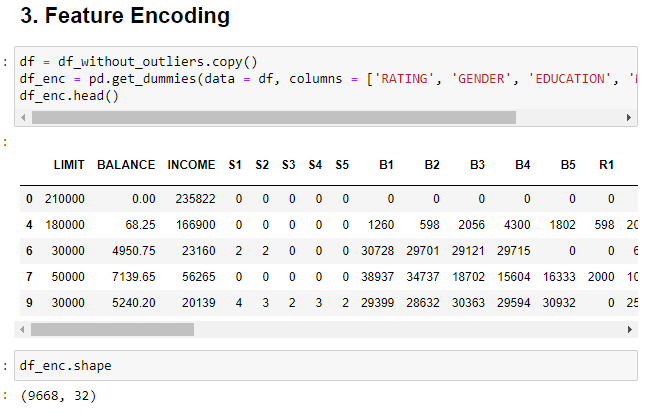


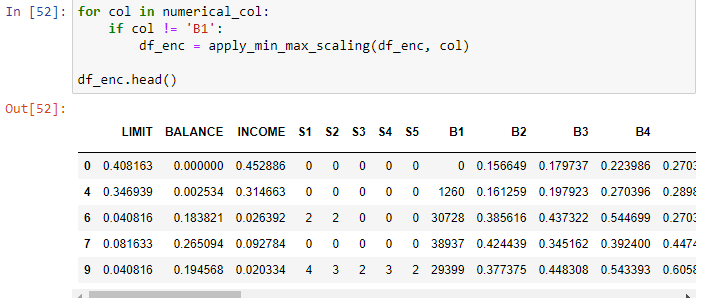




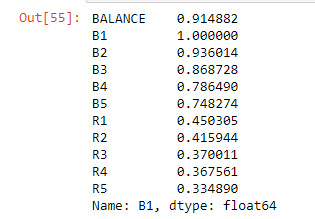
* **Step-4: Cleaning AGE column and convert it into a Categorical**
* There are -1 and 199 in the AGE column, it is an invalid age value, and those rows that have invalid age values should be dropped.
* Also AGE should be converted into a Categorical feature so that it can affect the target feature prediction well, as the result, age values were grouped into AGE BIN, such as “old”, ‘middle’ and “young”.
* 
* **Step-5: Remove Outliers**
* As there are many outliers in numerical columns, outliers can be removed before linear regression modeling.
* As a result, we could get 9668 rows from 18720 rows.



* **Step-6: Feature Encoding:**
* All categorical features except S1~S5 should be encoded before model predicting, because any python machine learning models are requiring only a numerical dataset, so it can be done by pandas.get\_dummies() function. 
* **Step-7: Data Scaling.**
* Data normalization and scaling is the important technique for Modeling, here target column should not be scaled.
* I applied min-max scaling for all numerical features(except one-hot encoded features and target feature-B1).



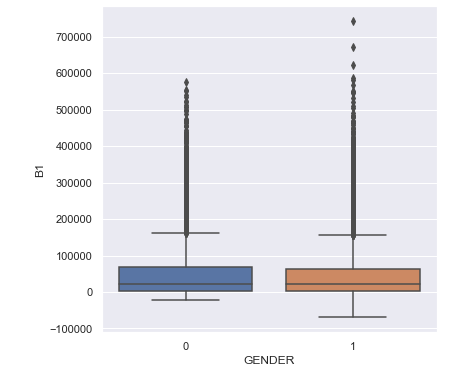
* **Step-8: Feature Selection**
* As there are many features and some features can affect negatively the target feature prediction, and it may take much time to train the model as there are many features, so I can apply the feature selection method before modeling
* I could get only 10 features that is highly correlated with target feature are As the result, most correlated features are 'BALANCE', 'B1', 'B2', 'B3', 'B4', 'B5', 'R1', 'R2', 'R3', 'R4', 'R5'.



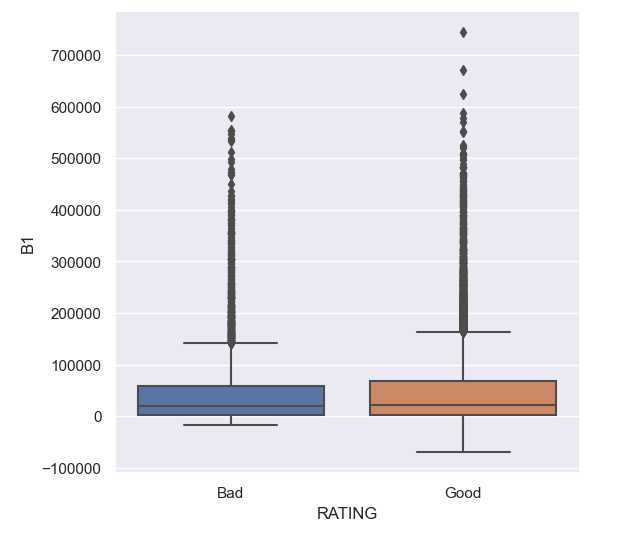
# Question 3:

Five (5) relevant insights of the data, with supporting visualization.

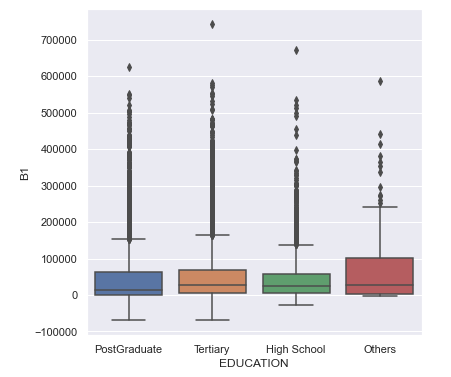
* **Insight-1:**
* There isn’t any meaningful difference between some categorical features and the target column(B1).
* Average B1 and distribution of B1 are almost the same for GENDER



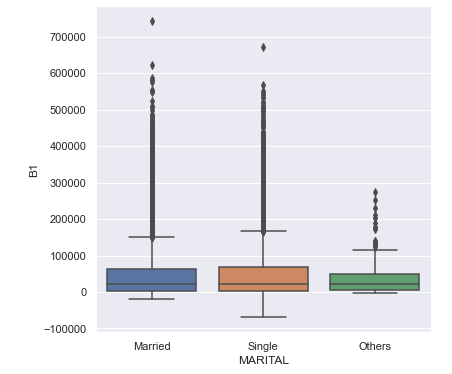
* Average B1 and distribution of B1 are almost the same for RATING.



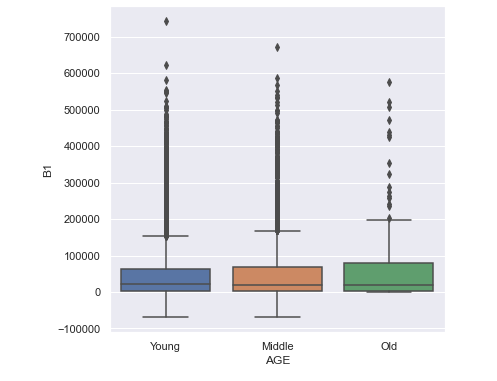
* Average B1 and distribution of B1 are almost same for EDUCATION.



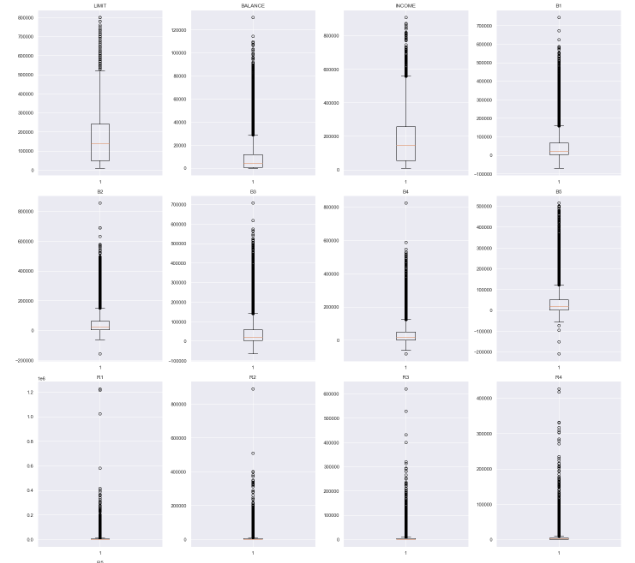
* Average B1 and distribution of B1 are almost same for MARITAL status.



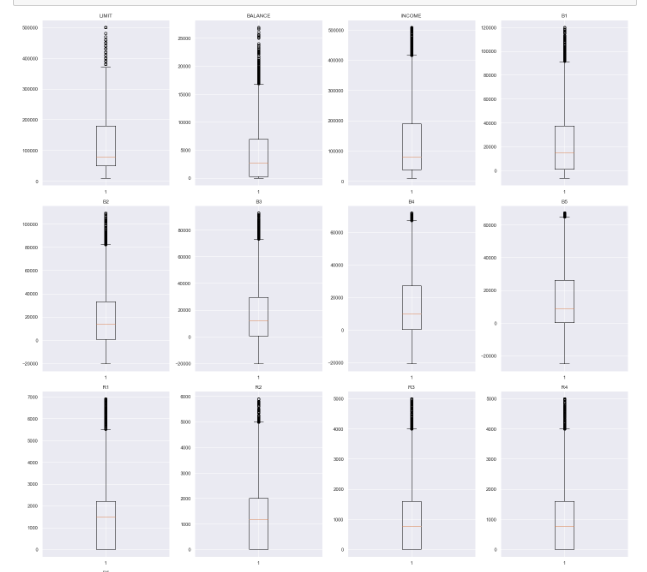
* After converting AGE into a grouped categorical feature, average B1 and distribution of B1 are almost the same for the AGE group.



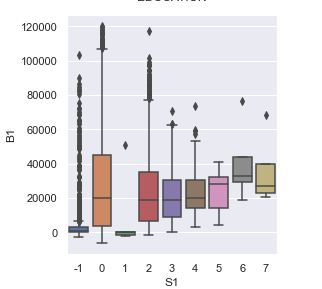
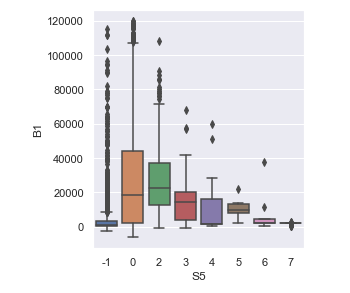
* As a result some categorical features(RATING, AGE, EDUCATION, MARITAL, GENDER) are not so important features for B1 predicting.
* **Insight-2:** There are many outliers for numerical features and it has been dropped by using IQR.
* Visualizations of outliers for original dataset(Dots are representing outliers).

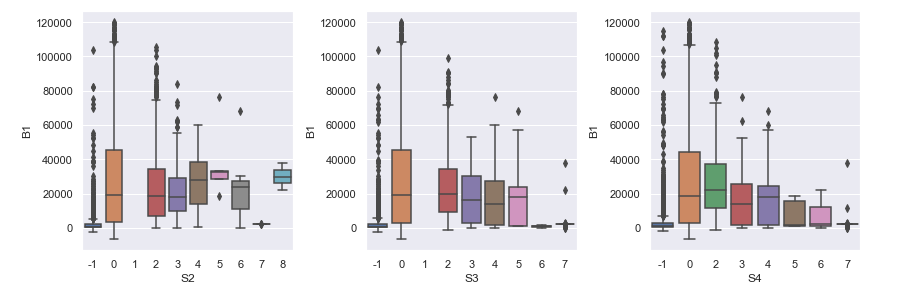


* Visualization after dropping outliers.

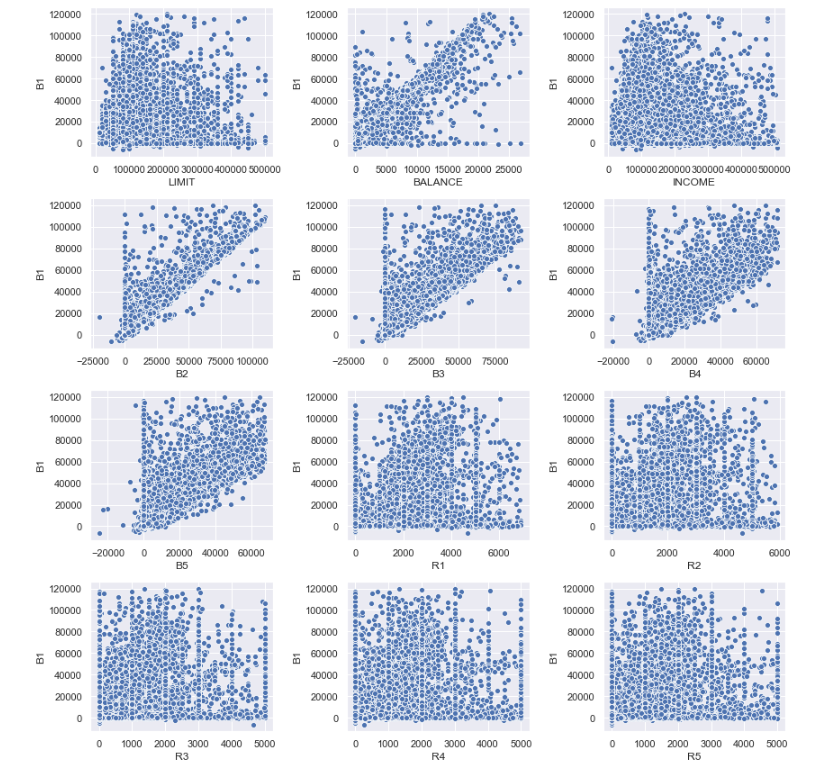


* **Insight-3**: After dropping outliers, some categorical features are highly correlated with target column(B1).
* S1 ~ S5 are highly correlated with the Target column(B1).

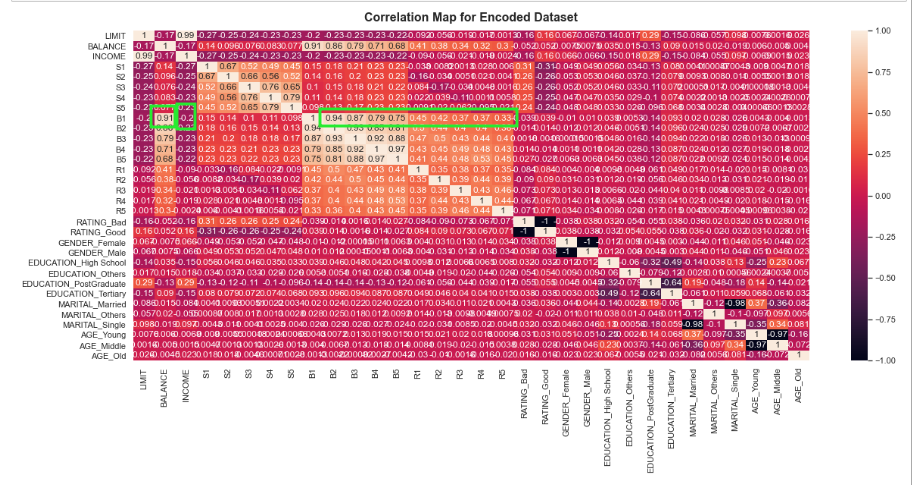
 



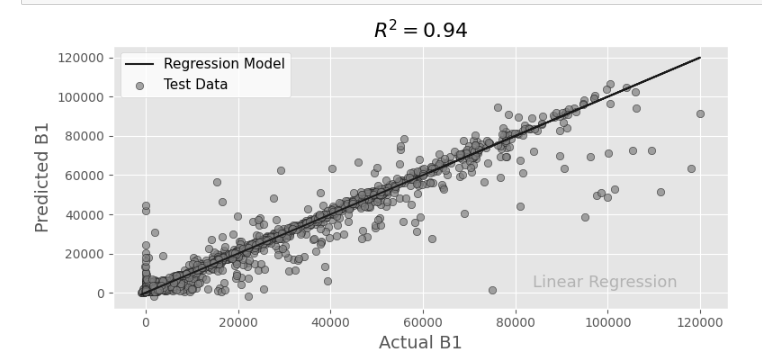
* Balance, B2, B3, B4, B5 are the most important features that correlated with target column(B1).



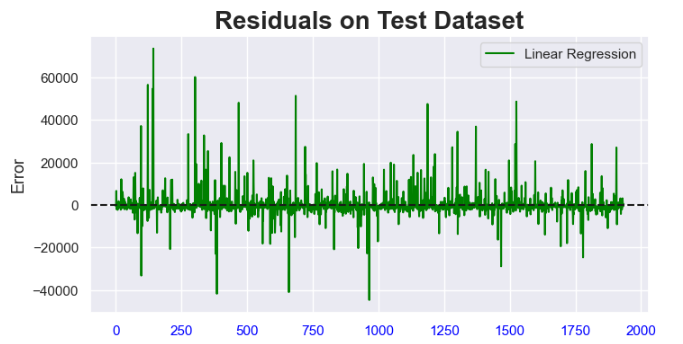
* **Insight-4:** Balance, Income, B2, B3, B4, B5 and R1, R2, R3, R4,R5 are important features that could predict target column very well, these features should be selected for Regression Modeling.



* **Insight-5**: Linear Regression Modeling result is so good to predict B1.
* This plot shows how the predicted results are similar to actual test values.



* This plot shows residual, it should be closed 0 when the actual and predicted value is the same.

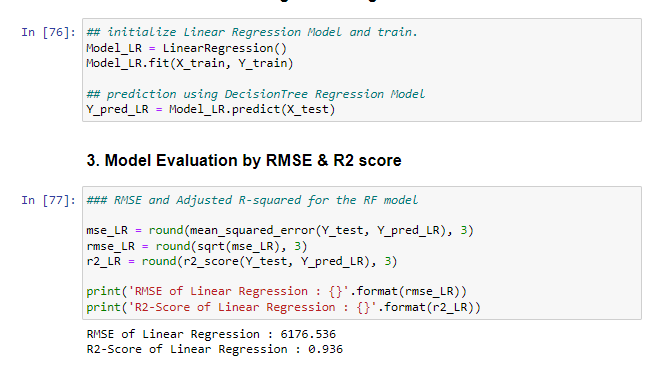


# Question 4:

* Regression is a form of machine learning algorithm that help to predict continuous value outputs. In Python, Linear regression help to find the relationship between two variables based on the dependent variable (x) and independent variable (y). It can help to predict the dependent variable (x) based on the independent variable (y).

In the given dataset, B1 is the customer billable amount in 1st month and it is the dependent variable(x). The value can change based on the Gender, Limit, and Age of a customer and we can use all this data to predict the billable amount of customers in the most recent month.

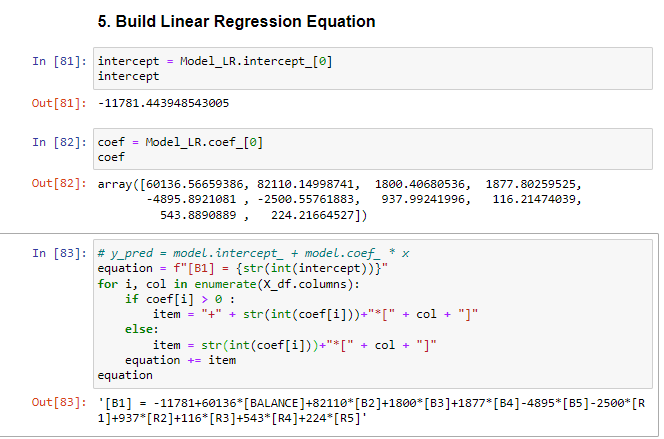
* Using LinearRegression() in Sklearn library, and RMSE, test predicting is 6176 and R2 score is 0.936 (=93.6%). Result is good.



# Question 5:

* From the prediction model result, we can get the regression equation to predict B1 from BALANCE, B2, B3, B4, B5 and R1, R2, R3, R4, and R5.
* Linear Regression Equation:

**[B1] = -11781+60136\*[BALANCE]+82110\*[B2] +1800\*[B3] +1877\*[B4]-4895\*[B5]-2500\*[R1]+937\*[R2]+116\*[R3]+543\*[R4]+224\*[R5]**



* **Key Insights:**
* BALANCE, B2, B3, B4, B5, and R1, R2, R3, R4, and R5 are the most important features to predict B1.
* It is said that customers' payment and refund policies over the past 4 months play a decisive role in predicting the current customer's ability to pay an amount.
* It can be seen that characteristics such as age, gender, educational background, and education level of the customer have little effect on the prediction.

**Question 6:**

Submitted Jupyter notebook