CCT College Dublin

Assessment Cover Page

|  |  |
| --- | --- |
| Module Title: | Strategic Thinking |
| Assessment Title: | Strategic Thinking Continuous Assessment 1 |
| Lecturer Name: | James Garza |
| Student Full Name: | Nick McNamara; Anna Georgieva; Octavio Rieu |
| Student Number: | [sbs23022@student.cct.ie](mailto:sbs23022@student.cct.ie);  sbs23039@student.cct.ie; sbs23024@student.cct.ie |
| Assessment Due Date: | 14/11/2023 @ 11.55pm |
| Date of Submission: | 14/11/2023 |

Nick McNamara

Anna Georgieva

Octavio Rieu

Declaration

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

Prediction for lending loans

Word count:

## Table of contents

[Table of contents 3](#_w2b5qc91shel)

[List of Figures 4](#_qu0qt0jsoamm)

[Introduction 5](#_37041gfy3b00)

[Business understanding 5](#_f27wby1pvcsx)

[Hypothesis 6](#_hnim1g2ex6fw)

[General goal 6](#_661oi4dtxqam)

[Success criteria/indicators 6](#_ibmk5ux239dh)

[Technologies 6](#_5poayr8hwgn)

[Libraries 6](#_e3jm6rp2sitl)

[Models 7](#_6ady5g2ljeva)

[Datasets and source 7](#_597mabpt9cn9)

[Descriptive statistics 11](#_sai3ymlwptfj)

[Data preparation 15](#_qekc2b3jdv3b)

[Model implementation 21](#_d7gklkhanrn7)

[Deployment 22](#_13tpygmpw3nl)

[Challenges 23](#_y7n4pc94gqui)

[Results and analysis 23](#_mab4aegwkzyj)

[Conclusion 25](#_nsqnf26gn17b)

[Reference list 26](#_3z44o5wi77sr)

## List of Figures

[Figure 1: Libraries and models 7](#_xrqck0vrnn1s)

[Figure 2: Read CSV 8](#_himytwo15w93)

[Figure 3: See all the attributes 8](#_z3di9jcynqes)

[Figure 4: Check null values 9](#_ycldsb202ld3)

[Figure 5: Totals of null values 9](#_t3sefk9nauhh)

[Figure 6: Remove null values 10](#_4yu8waw9xi69)

[Figure 7: Check data types 10](#_52bekna2rd2p)

[Figure 8: Descriptive table of datasets 11](#_h9jh7dfiq5ym)

[Figure 9: Percentage of approved and denied loan 11](#_khvemvc7uvug)

[Figure 10: Females vs Males income 12](#_bx16pxf28hjs)

[Figure 11: Loan status vs Education 12](#_nctg842ftxzn)

[Figure 12: Loan status vs Married 13](#_kycv42dx77q5)

[Figure 13: Loan amount frequency 14](#_ak5qcbt0uq9r)

[Figure 14: Loan amount frequency vs Loan Status 14](#_875zgd35vnuy)

[Figure 15: Loan status by Dependents 15](#_wv67ug7wdml5)

[Figure 16: Data types for the datasets 15](#_q2ynb1m8t9sb)

[Figure 17: Data normalization 16](#_hromfdg7dhm4)

[Figure 18: Drop independent variable from the dataset 16](#_gfobsb34hn79)

[Figure 19: Libraries and models 17](#_a7t0s45qtjqb)

[Figure 20: Check datatypes and change object data to float 18](#_nzdyvc82v180)

[Figure 21: Frequency of all variables 19](#_u0wsqbrdnrcg)

[Figure 22: Heatmap for all variables and correlations 19](#_uf27cmjympa1)

[Figure 23: Association between variables 20](#_ev0dju3z9b8r)

[Figure 24: Concad method 21](#_oao0287p03e0)

[Figure 25: Split dependent variable and independent variables 21](#_t5xcf3u9amhy)

[Figure 26: Decision tree classification 21](#_dyuwjhj64mfy)

[Figure 27: Linear regression, mean squared error 22](#_dpzwmdm5ejb1)

[Figure 28: Random forest regressor 22](#_4csz0jcmd5kf)

[Figure 29: Saving Linear regression model in .pkl format 22](#_281wdnpttb15)

[Figure 30: Linear regression deployed model 23](#_aq5w8durvhgb)

[Figure 31: Coefficient magnitudes 24](#_q2fjlispjprj)

[Figure 32: Additional features 24](#_50hyjtme80ik)

[Figure 33: Ascending order of coefficients magnitudes 25](#_x7il7d13iq83)

## Introduction

Word count: 998

This semester, our team should conduct a comprehensive assessment of the capstone project from Semester One, refining its strengths and identifying areas for improvement. Equipped with the insights and expertise gained from successfully completion of Class Assignment One in the first semester, combined with the advanced knowledge we have gathered throughout the second semester, we are in position to enhance and augment the project. Our approach will involve the deployment of at least three machine learning models for two datasets and optimise their performance through hyperparameter tuning. We will validate the outcomes to ensure the robustness of our results. We are committed to advancing the project. Our efforts from the last semester were foundational, setting the stage with a well-crafted hypothesis and a suite of strategic questions, leading us toward meaningful solutions. The current semester's objective is to solidify these initial findings, ensuring that our position is both well-supported and resilient. There is mutual agreement within our team to continue our focus on loan defaulting within the financial services sector, a subject that remains exceedingly relevant and compelling.

The recent surge in the Central Bank's interest rates has triggered a domino effect, prompting local Irish banks to implement corresponding increases in their rates. This economic instability and inflation make our project not only timely but critical, providing a rich context for our continued exploration and analysis. According to the article of (O’Halloran, 2023), increasing interest rates have unsettled prospective homeowners, with a majority now opting for fixed-rate mortgages. The research, released by Myhome.ie, a property portal owned by The Irish Times, indicates that surging borrowing costs along with worries about housing availability are putting pressure on the real estate market. The survey reveals that *66%* of aspiring homebuyers are concerned about how escalating interest rates might impact their home purchasing capabilities, a significant increase of *20%* from the previous survey in March. Myhome.ie reports that the climb in interest rates has alarmed mortgage seekers, noting that over half of them are in search of mortgage agreements extending beyond five years, and three-quarters are intent on obtaining a fixed-rate mortgage. In response to the change of ECB, (Healy, 2023) *Bank of Ireland*, for example, has increased its variable mortgage rates by *0.25%* from 27th of October this year, whilst Allied Irish Banks(AIB) increased its variable interest rates for mortgages by *0.55%* for Private Dwelling Homes(PDH) since 14th August, 2023 (AIB, 2023).



#### Figure 1: AIB PDH Variable Rates

Regardless of the gloomy forecast for mortgage lending, it appears that first-time buyers purchasing power remains solid, no matter that the mortgage activity slowed down in Q2 (Murphy, 2023). The drawdown value is €284,397 which is at its peak since 2003 and there are a total of 9,896 new approved mortgages. The number of mortgage loans dropped by 5.7% and by value with 3.6% in comparison to Q1, 2023.

We attempted to find datasets that represent a sample of loan borrowers for the Irish market, especially data, which is recent and actual with the current mortgage climate, however it has proven difficult as there are no public repositories available for use. If we did manage to find this kind of dataset, it would have been in juxtaposition with our hypothesis, because our findings supported the theory that loans are granted to borrowers, who have prior credit history and track record of a stable repayment capability. A relevant question to consider is whether the evaluation process for loan applications of ‘Dream Housing Finance Company’ (Kaggle datasets, 2023) is consistent with the standards applied to citizens within Ireland.

In the first semester, we examined the research conducted by Lee and Lee (2018), detailed in their book which asserts that credit history is the primary consideration in lending decisions for a diverse array of borrowers, including individuals, partnerships, corporations, clubs, societies, and trusts. Lee and Lee describe that a credit score is derived from a person's past and current credit and this can predict their likelihood of honouring debt commitments. Lenders leverage these scores to assess the viability of a loan applicant, to decide on the interest rates, and to determine the extent of credit limits. The credit score is influenced by several factors: payment history on various accounts such as credit cards and mortgages accounts for *35%* of the score; legal and financial blemishes like lawsuits, bankruptcies, and court judgments make up *30%*; and the effect of opening multiple new credit accounts suggests a negative impact, as it might signal repayment challenges, contributing to *15%* of the score. Our capstone project hypothesis corresponded with the conclusions drawn by Lee and Lee (2018).

Nevertheless, considering the statistical overview for mortgage lending in Ireland, can the findings of Lee and Lee (2018) be considered universally applicable and serve as foundational principles across diverse banking systems? Does the discretion in approving loans depend on the economic climate of the country or is it tailored to the financial circumstances of individual applicants? We will evaluate all datasets features again and determine what factors weigh in these decisions. In addition to analysing Dream Housing Finance Company datasets, we will also explore the 'Loan Default Prediction' dataset from Kaggle. This particular dataset allows us to confront a critical and industry-relevant machine learning challenge—the prediction of loan defaults. It encompasses a distinctive set of 255,347 records across 18 different attributes, providing a rich testing ground for enhancing our predictive modelling skills. By examining both datasets, we aim to further validate our hypothesis for a second time. The newly included datasets have features, which are lacking in the previous datasets and are quite interesting to analyse, such as ‘*CreditScore*’, ‘*Months Employed’*, ‘*NumCreditLines’*(the number of credit lines the borrower has opened), ‘*InterestRate’*(a feature which is essential when evaluating the decision to lend a loan to an individual), ‘*DTIRatio’*(Debt-to-income ratio, which indicates the person’s debt in comparison to their income), *HasMortgage*, ‘*HasCoSigner’*(whether the loan has co-sign) and lastly, but not the least by importance *‘Default’*.

## Business understanding

Word count: 1365 up to page 12

We have used Cross-Industry Standard Process for Data Mining (CRISP) methodology for this project and our main objective is to reduce the risk of lending to borrowers who are likely to default on their loan.

The first dataset consists of borrowers whose loan status is approved or rejected and the second one consists of borrowers, who defaulted on their loan or not. We will evaluate all the variables in the datasets and determine which ones have the highest impact on the decision to approve or decline a loan. The data mining process will involve data preparation, such as cleaning the datasets from NaN or missing values, data normalisation, analyse the relationships between the variables, train and test the datasets with a couple of models such as Decision Tree Classifier, Linear Regression, Random Forest Classifier, neighbours Classifier and based on the results make a conclusion, which model is the best performing one. In order to produce this project plan, we have used a Ghantt chart with a timeline on when each task should be completed by.

### Hypothesis

For this semester, our focus will be on analysing two distinct datasets. The first dataset contains loan applications with their respective approval statuses, which seemed to be influenced by the applicants' Credit History. The second dataset offers insight into which loan applications resulted in defaults. Our objective is to assess the potential risk faced by banks when approving loans to clients. Through our analysis, we aim to validate the hypothesis that the Credit History attribute is the most significant predictor in determining loan approval outcomes.

### General goal

Our primary path is to minimise biases and maintain objectivity in our handling and analysis of data throughout the duration of this project.

### Success criteria/indicators

Success of this project would be determined by finding a correlation between the individual customer circumstances and a proclivity to default on a loan. To achieve the best predictions and results with higher accuracy, we will use machine learning models and algorithms such as linear regression, random forest algorithms and others to predict the outcome. We would hope one of these algorithms would provide us with test results in excess of 98%.

### Technologies

The programming language of choice is Python. We use Jupyter Notebook in order to work on the datasets. For the project, version control of CA2 is required, hence we installed Github Desktop.

### Libraries

We use the following libraries:

* Numpy
* Pandas
* Matplotlib.pyplot
* Seaborn

*For preprocessing data we use:*

* Label Encoder
* MinMaxScaler
* StandardScaler

*Training and testing the data:*

* import train\_test\_split

*Evaluating the model:*

* Metrics
* Accuracy\_score

### Models

* Decision Tree Classifier
* Linear Regression
* Random Forest regressor

### Datasets and source

The first dataset, which we use for this project is called ‘*Home Loan Approval’* and the source is from Kaggle (Konapure, 2023). This dataset is owned by a finance company, which lends loans to people who want to buy properties in rural, semi-urban and urban area. They need to automate the approval process by segmenting the customers’ eligibility. The second dataset called ‘*Loan Default Prediction Dataset’* is also from Kaggle, however it contains data of individuals who defaulted on their loans (NIKHIL, 2023).

**Cleaning and preparing ‘*Home Loan Approval’(CA2 Jupiter Notebook)***

### Figure 2: Libraries and models

Pandas package for Python was used to analyse and handle the datasets. This library is suitable and efficient in handling data (Müller and Guido, 2017). NumPy is the numerical package for Python, is also utilised. We select the seaborn library for data visualisation, which generates fascinating and practical statistical visualisations like heatmaps, bar charts, pie charts, scatter plots, and others. As an alternate library for producing high-quality graphs and charts, Matplotlib is also imported.

We have two types of datasets belonging to ‘Home Loan Approval’. One CSV is used to train data and one CSV is used to test data, so we can evaluate the success of the trained data.

### Figure 3: Read CSV

We use loan = pd.read\_csv(‘loan\_train.csv) code and load the datasets into pandas dataframe and we name it ‘loan’. The CSV files are in the same GitHub Desktop directory as the Jupyter notebook, which contains this code.

We run loan.head() to display the first 5 rows and all columns, so we can quickly check what kind of data we work with. The attributes in this dataset are categorical Loan\_ID, Gender, Married, Dependents, Education, Self\_Employed, Credit\_History, Property\_Area, Loan\_Status and numerical attributes - ApplicantIncome, CoapplicantIncome, LoanAmount and Loan\_Amount\_Term.

### Figure 4: See all the attributes

It is a common issue to handle missing values in a dataset and we will use the **isnull** method in order to identify them (Harrison, 2019).

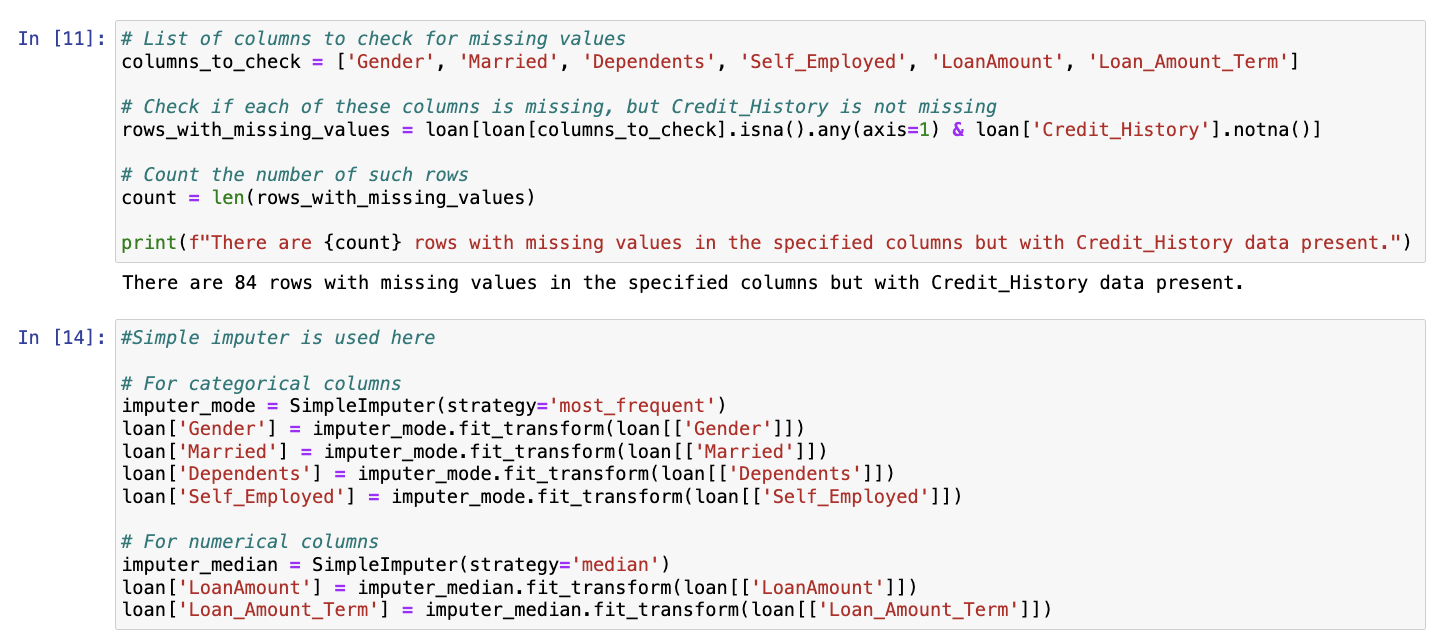
### Figure 5: Check null values

This dataset is not a high-dimensional one as we have only 11 attributes to analyse and 614 observations. We should exclude Loan status, because it is a dependent variable in this data mining process and Loan ID, because it doesn’t have impact on the analysis and it is an identifier of the loan application. The data type of ‘*ApplicantIncome*’ is int64 and it is converted to float64.

We checked the count of missing values.

### Figure 6: Totals of null values

In the previous semester we dropped any row which contained a missing value. For this capstone project, it has been decided that instead of dropping all unknown values, it will be a better approach to impute all of them, because this could impact the overall results with such a small dimensional dataset. It is essential to understand the importance of these features and the approach is taken accordingly. For example, at the beginning of this report, we mentioned that *‘Credit History’* is the most important influencer on credit loan approvals. There are 84 rows with missing values in the specified columns: *‘Gender’*, *‘Married*’, ‘*Dependents’*, ‘*Self\_Employed’*, *‘LoanAmount’,* *‘Loan\_Amount\_Term’*, but with ‘*Credit\_History’* data present. If the unknown values are dropped from all these features, we can impact significantly on the machine learning models outcomes. This is the reason why, by employing SimpleImputer, the most frequent values were used for the categorical features, such as ‘*Gender’,* *‘Married’*, *‘Dependents’* and *‘Self\_Employed’* and the mean or median for numerical data - *'LoanAmount'* and *'Loan\_Amount\_Term'.* The sklearn.impute.SimpleImputer in scikit-learn provides a convenient way to impute missing values in a dataset. By default, for numerical data, it can replace missing values using the mean or median of each column, while for categorical data, it can use the most frequent value (mode). As of sklearn's documentation in 2023, these strategies are widely applied due to their simplicity and effectiveness in many scenarios.



### Figure 7: Imputing categorical and numerical values

However, for a critical feature like *'Credit History'* that has 50 missing entries, a more sophisticated approach may be necessary. Given the feature's significance, a multivariate imputation method could be more appropriate. This would involve using the entire dataset and employing all available features to estimate the missing *'Credit History'* values, thus preserving the underlying data structure and relationships (scikit-learn, 2022). In the section for 6.4.4. Nearest neighbours imputation article, it states that the KNNImputer class utilises the k-Nearest Neighbors technique to input missing values in a dataset. It employs the 'euclidean\_distances' metric by default to estimate the closest neighbours, even when there are missing values. For imputing a specific feature, the algorithm considers values from the nearest 'neighbours' that have non-missing values for that feature. These neighbour values are then averaged, either equally or weighted by their distance, to fill in the missing entry. If multiple features are missing from a sample, the set of neighbours used for imputation may vary for each feature. In situations where the available neighbours are fewer than 'n\_neighbors,' the overall mean of the feature across the training set is used. However, if there is at least one neighbour within a certain distance, the average of these neighbours are either weighted or not and then is used for imputation. Features that are consistently missing across the training data are excluded in the transformation process.



### Figure 8: Imputing *‘Credit history’* with KNNImputer

Upon reviewing 'Figure 9', it becomes evident that in our prior project, we employed dimensionality reduction techniques on a dataset and removed 134 observations. This method, however, may not represent the most accurate approach to managing the missing values present within the dataset. Our analysis reveals that out of these observations, 84 rows are with missing values across various features, with the exception of 'Credit History,' where complete data exists for the corresponding rows. This suggests the need for a more careful strategy in data handling to ensure the integrity and utilisation of the dataset.

### Figure 9: Null values are removed - an approach used in the previous capstone project

### Figure 10: Check data types

### Figure 8: Descriptive table of datasets

loan.describe() method provides quick statistical data for the numerical columns. We need to check approved loans vs rejected loans and 69.2% of the loan applications are approved, whereas 30.8% of the applications are rejected.

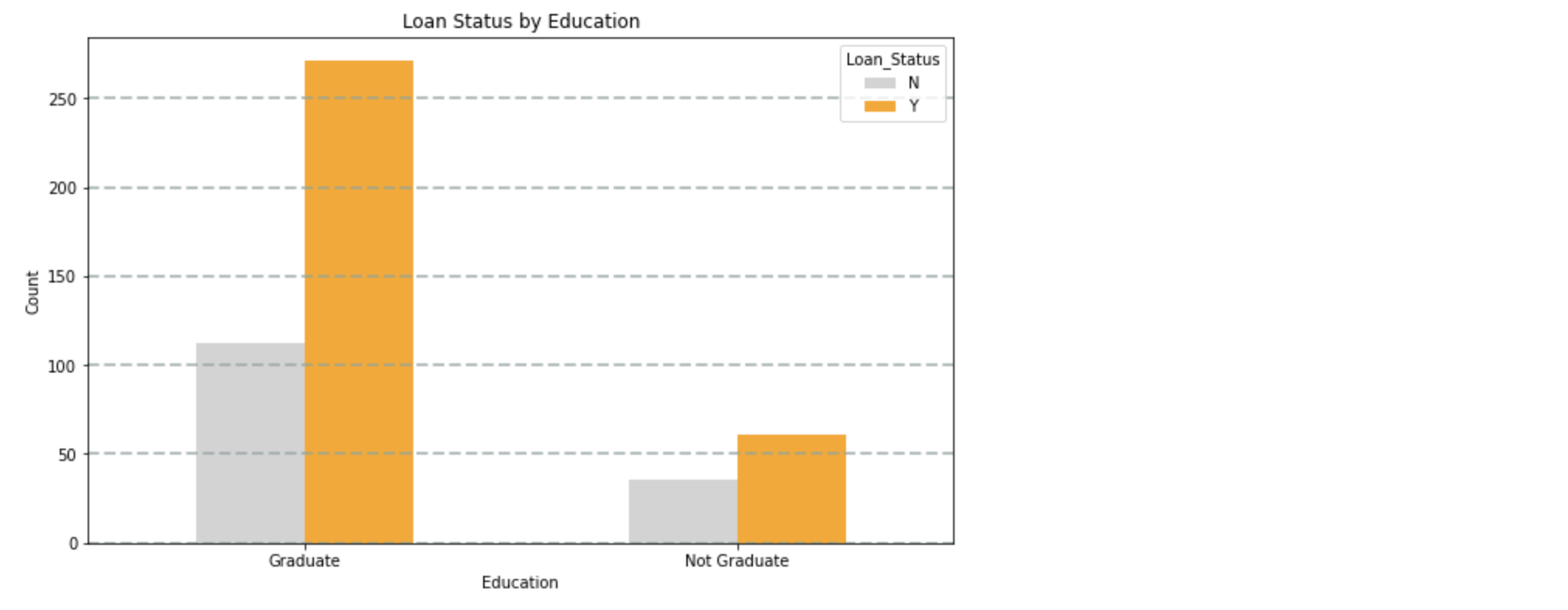
### Figure 9: Percentage of approved and denied loan

We used matplotlib library in order to visualize the above chart where we have done comparison of Males' vs Females' income. The chart shows that males have a better income than females.

### Figure 10: Females vs Males income

We created the same chart and compared two variables: 'Loan status' vs 'Education'.





### Figure 11: Loan status vs Education

The above data is with categorical variables ('Education') and ('Loan Status').We compare the two variables, so we can see that there is an apparent relationship between the education vs approval of the loan application. The same observations are with married couples, where we can see that Married couples are more likely to receive loan approval.



### Figure 12: Loan status vs Married



### Figure 13: Loan amount frequency

We are examining the data distribution of the variable 'Loan amount' in order to find out how frequently each value of this variable occurs and we have also included a piece of code which gives us an idea of what is the minimum, median and maximum value.

We also created a swarm plot, which shows that there are a lot of approved application for the lower loan amounts from the range of 100k to 200k.

### Figure 14: Loan amount frequency vs Loan Status

We compared the variable ‘Dependents’ vs ‘Loan status’. More than 175 approved loan applications belong to individuals with no children.



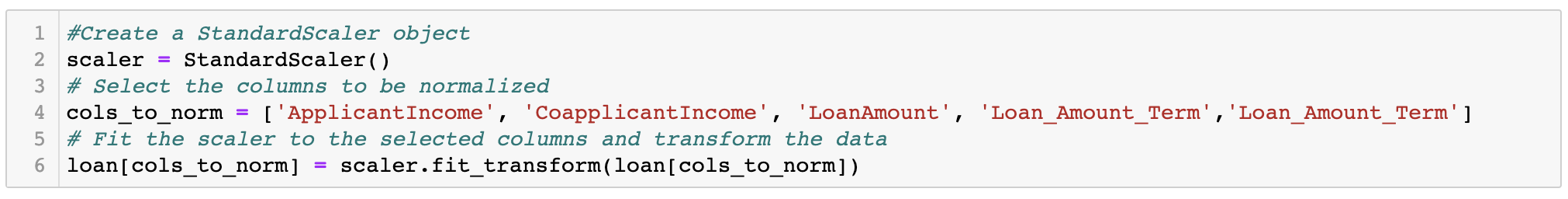
### Figure 15: Loan status by Dependents

## Data preparation

### Figure 16: Data types for the datasets

The above code displays numerical and categorical columns in this dataset.

We need to run Lable Encoder from sklearn libary in order to convert categorical string data into integer data. All the values are converted to zeros and ones - ‘Yes’ = 1, ‘No’ = 0, Property\_Area with values ‘Rural’ = 0, ‘Semi-urban’=1, ‘Urban’=2.

Figure 16: Data normalization with Standard scaler

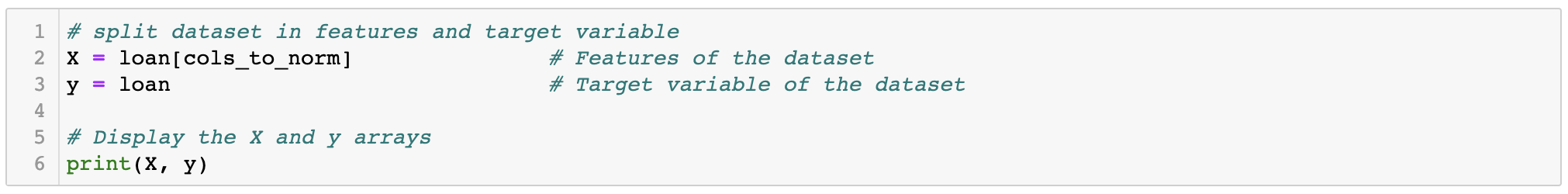
### Figure 17: Data normalization

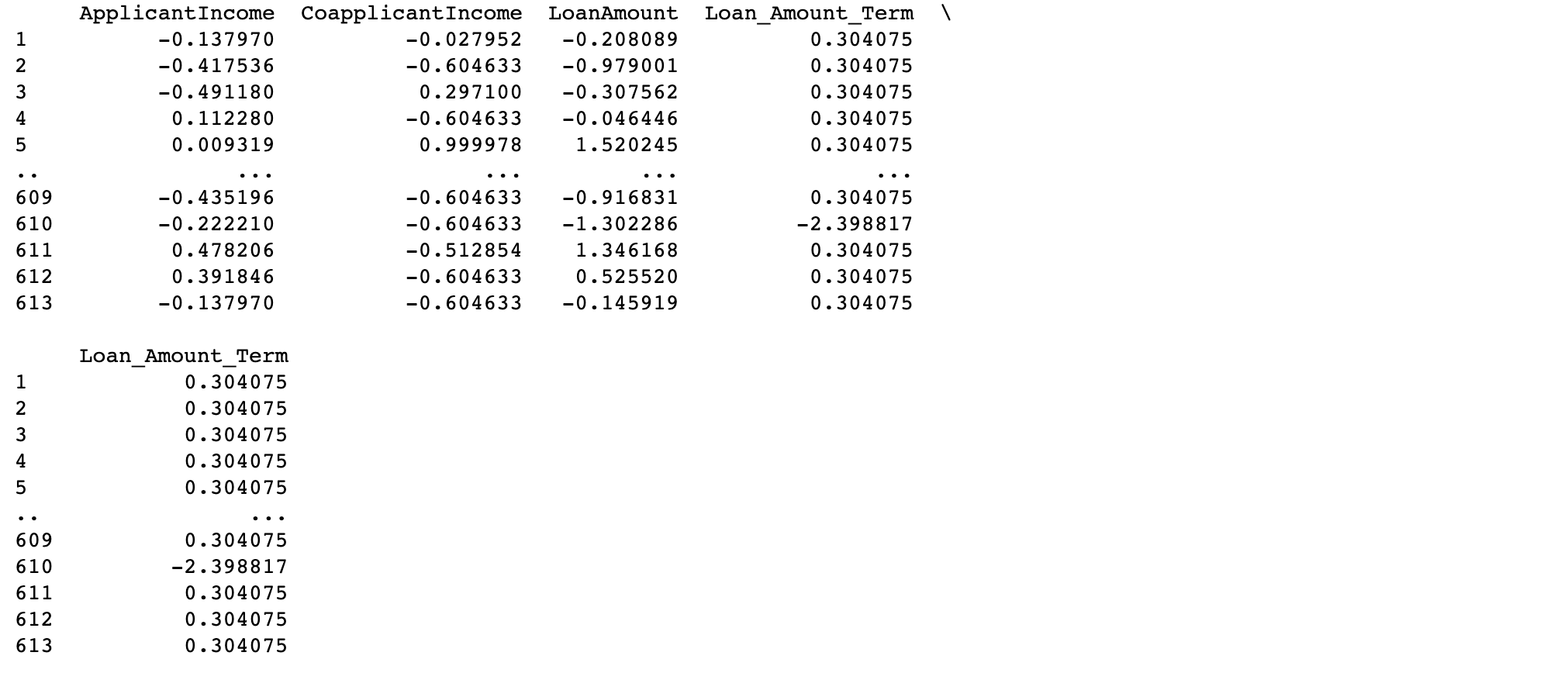
A crucial step in pre-prosessing data for a machine learning model is feature scaling. Since machine learning model sees only numbers, if the spectrum of numbers is quite broad from tens to thousands, the assumption will be that the higher ranging numbers have a superiority (Roy, 2020). This is why we apply it across the columns Applicant Income, Coapplicant Income and Loan amount.

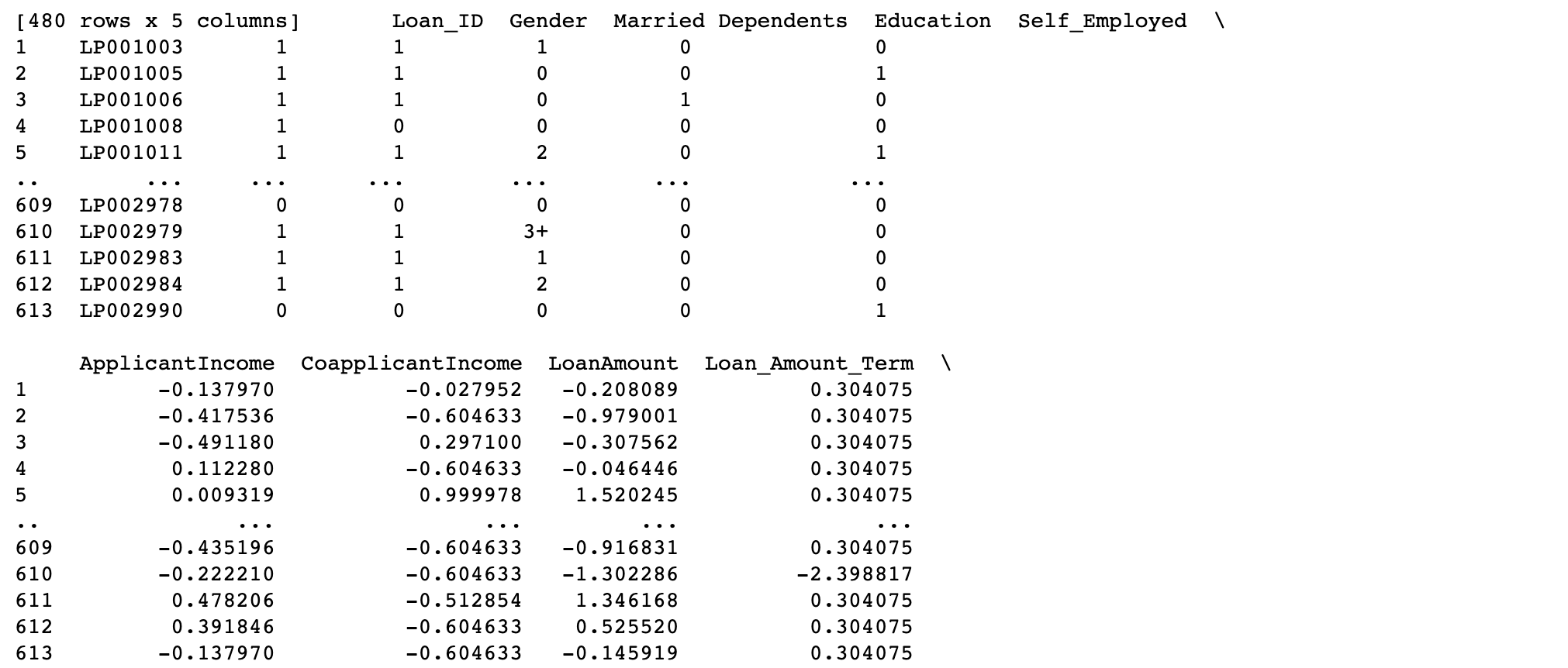
### Figure 18: Drop independent variable from the dataset

In order to evaluate the machine learning performance, we use the X and y as dependent and independent variables. The model will identify the relationship between the features in X and the target variable.

In addition to the above code, we need to define the X dependent variables as cols\_to\_norm.

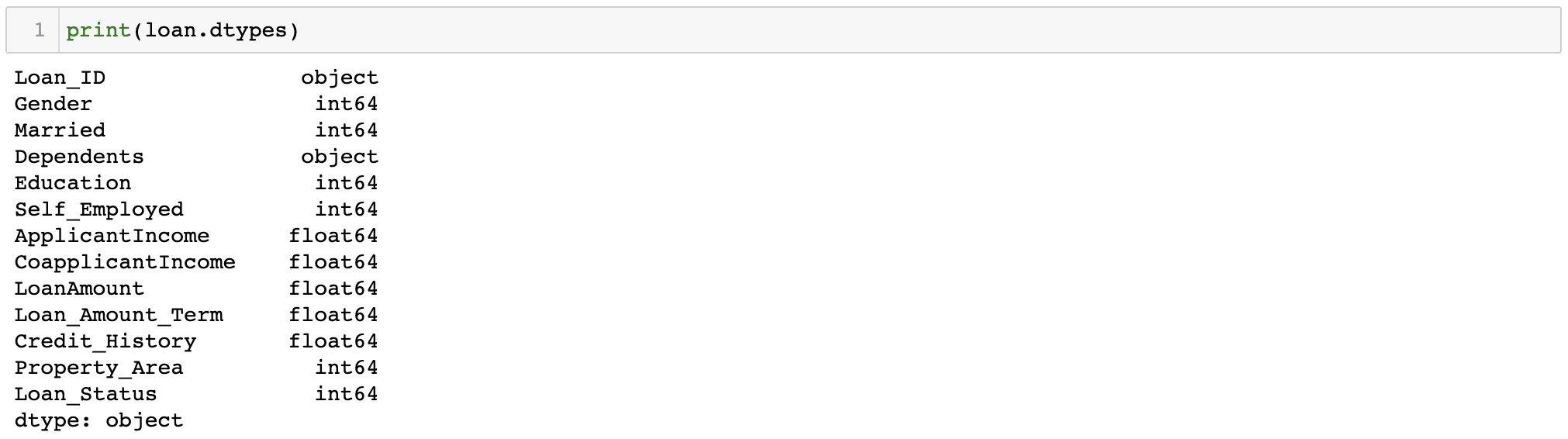






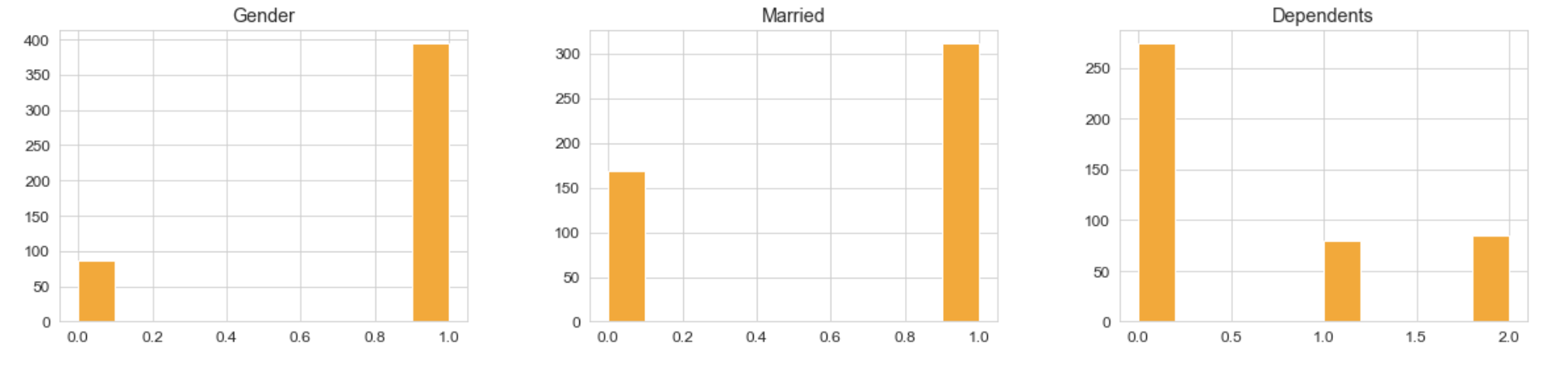
### Figure 19: Libraries and models

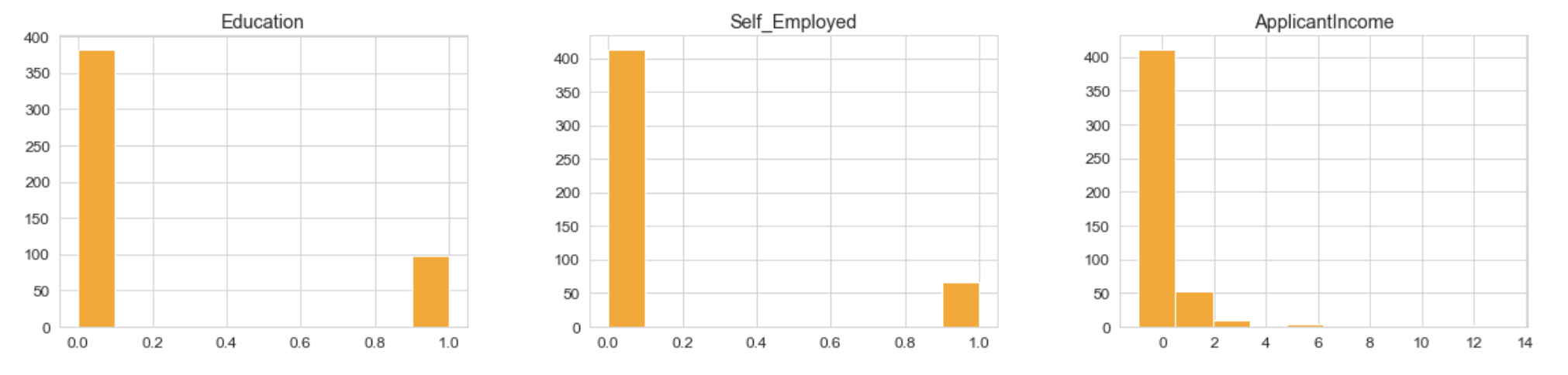
After checking again the data types, there is one more column to convert and it is ‘Dependents’.

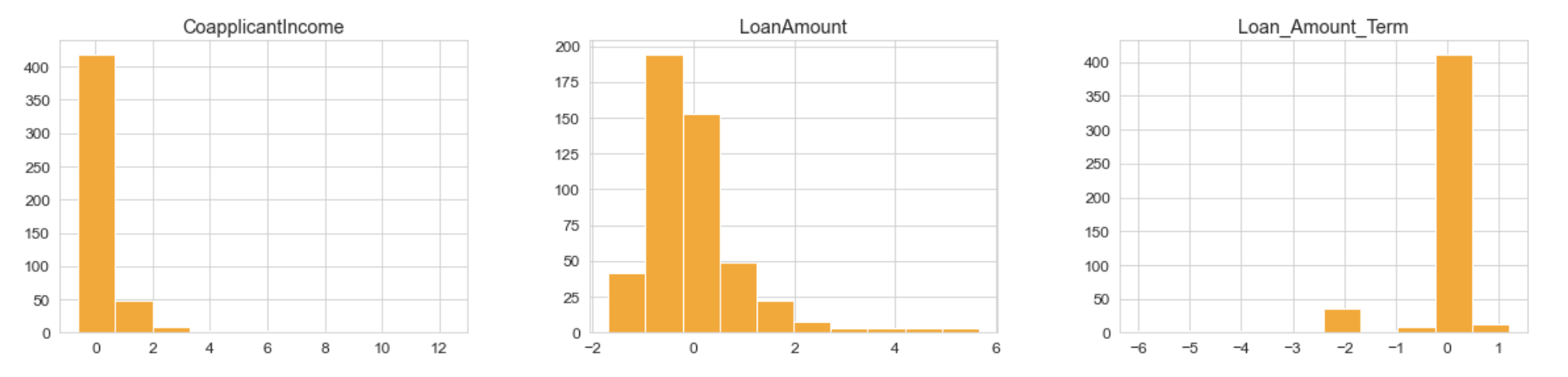


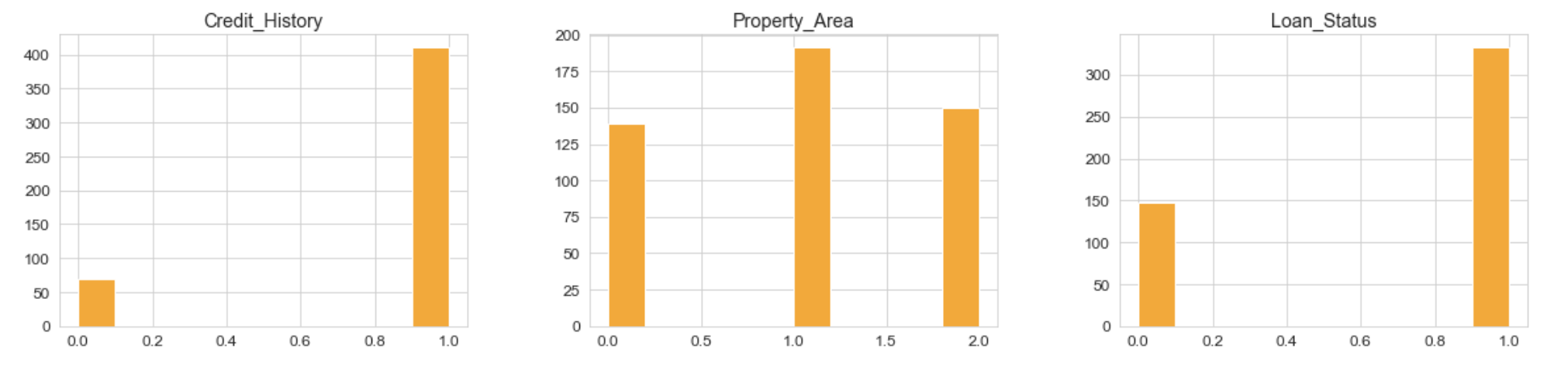
### Figure 20: Check datatypes and change object data to float

Since the data is normalised and cleaned for missing values it is essential to check the frequence of all variables.









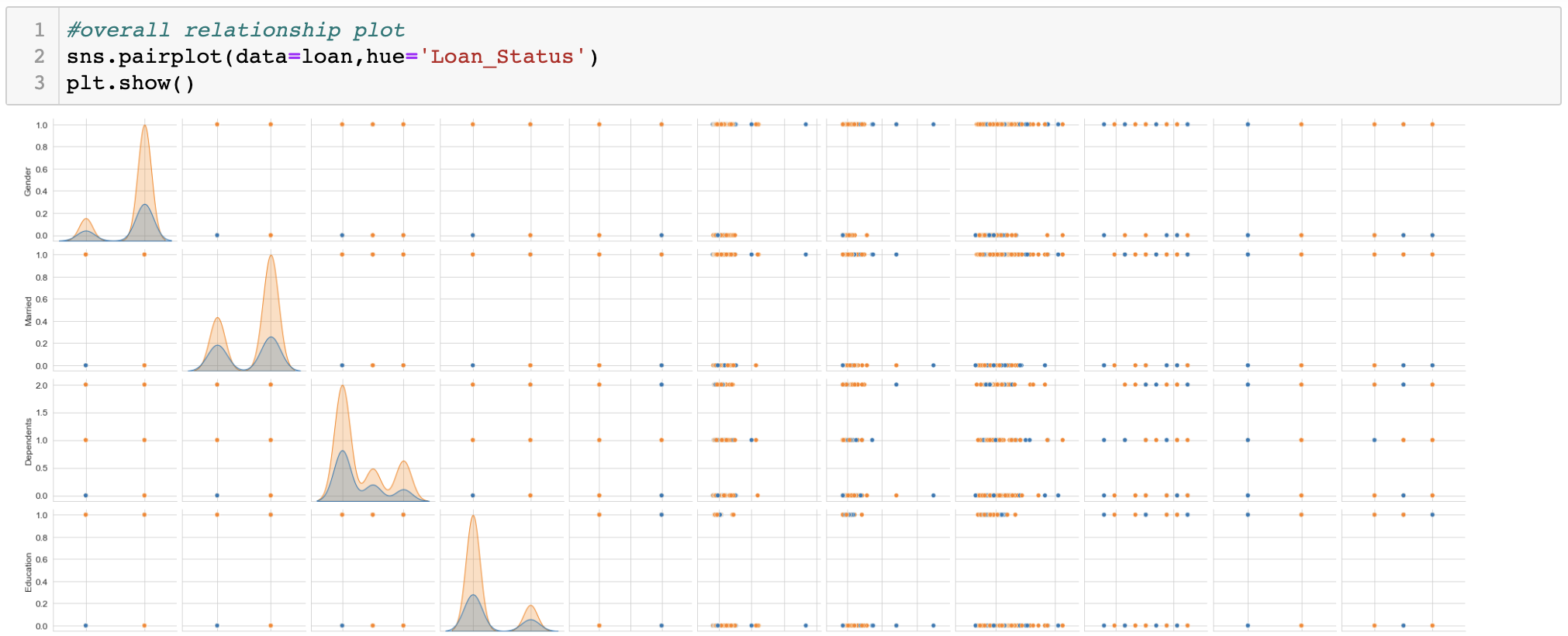
### Figure 21: Frequency of all variables

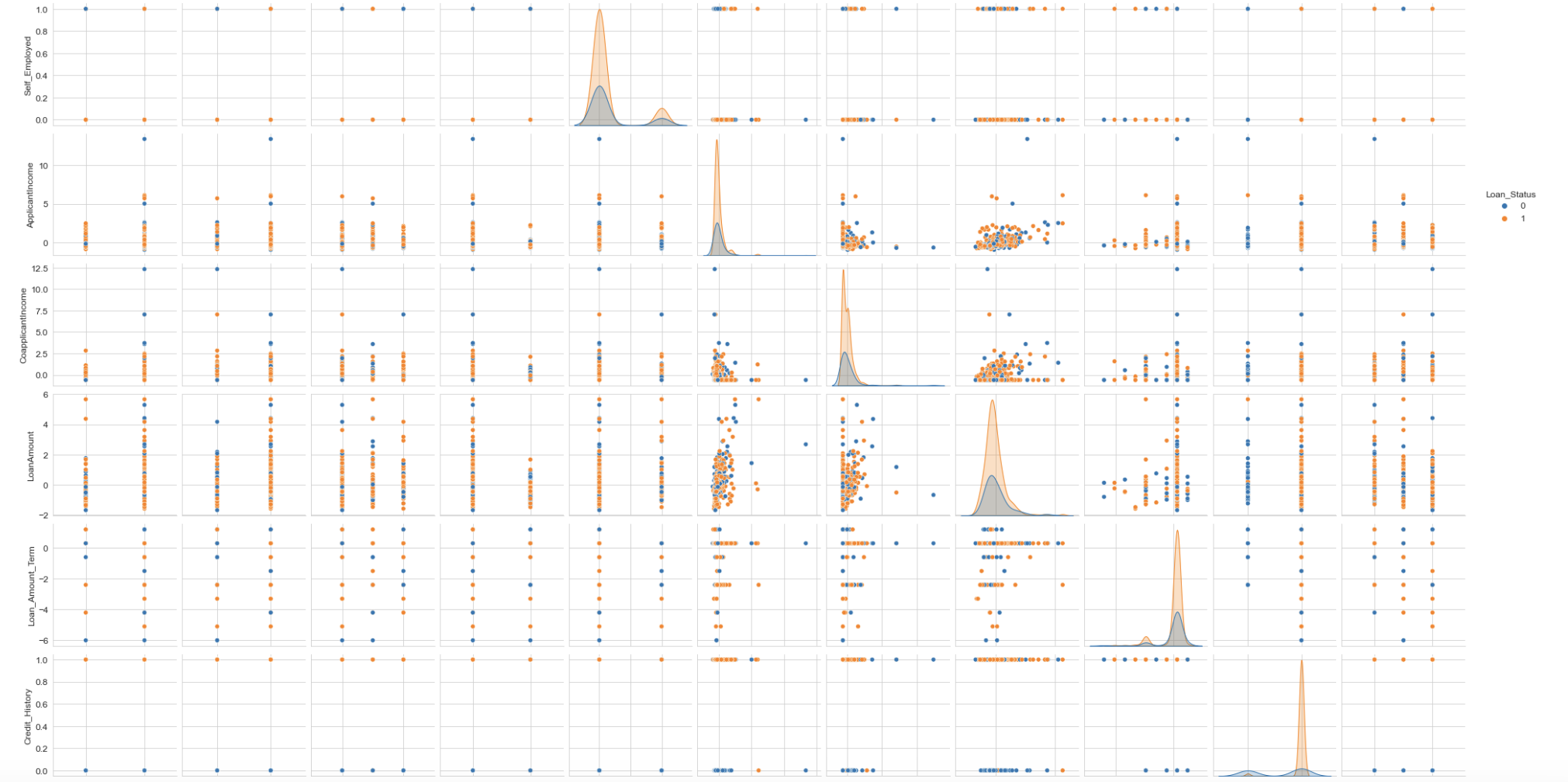
We run the following code to identify the correlations between variables.

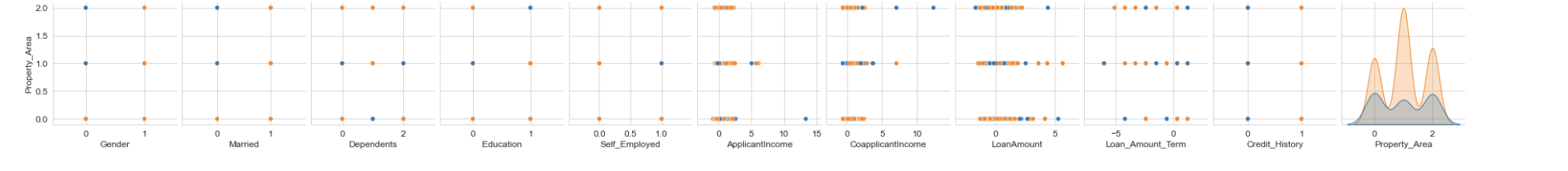
### Figure 22: Heatmap for all variables and correlations

The heatmap shows that there is 53% correlation between Loan\_status and Credit\_history and 52% correlation between ApplicantIncome and LoanAmount.

The below plot diagrams are produced with Seaborn library. The variables are coloured according to the target variable Loan\_status. The resulting plot displays all the relationships between variables in the dataset ‘loan’.

1





### Figure 23: Association between variables

By analysing the scatter plot and histograms, we can identify that there is a strong positive correlation between the approval of the loan amount and the credit history. The scatter plots showcase that males receive more loan approvals than females. Married couples get more approved applications than single individuals. If there is a mortgage application, the property area matters, as there are more approvals for semi-urban areas. Applicants who don’t have children or are not self-employed are more likely to receive loan approval. The most strongly associated independent variables to the target variable are ‘ApplicantIncome’, ‘Loan amount’, ‘CoApplicantIncome’ and ‘Credit\_History’. These numerical variables are appropriate for prediction of the independent variable ‘y’ = ‘Loan\_Status’.

## Model implementation

When we start splitting the data into training and testing and then run the code for DecisionTreeClassifier, we get errors that our data still contains NAN or missing values. We run pandas ‘concad’ method by combining the dependent and independent variables back again and after cleaning the data, the number of rows reduce to 439.

### Figure 24: Concad method

### Figure 25: Split dependent variable and independent variables

Our independent variables are Credit History, Applicant Income, Coapplicant Income, Loan Amount, Loan amount term. Dependent variable is Loan status.

We split the data into ‘train’ and ‘test’ again. We run the code for Decision Tree Classifier and we receive model accuracy 0.71 score.

### Figure 26: Decision tree classification

This accuracy score is not good enough, because this would mean that 29% of the applications could be misqualified. Money could be lended to applicants, who can not afford paying back the loan and this could be a financial risk for the bank. We did a test with just four features: 'Married', 'Education', 'LoanAmount' and 'Credit\_History'. The percentage went down to 60%.

We also run code for Linear regression model and we get a score of 0.14 mean squared error.

### Figure 27: Linear regression, mean squared error

The lower the value of mean squared error is, the better. The result is satisfying. We test other models to see if we get a better score.

### Figure 28: Random forest regressor

Random Forest Regressor model is with high mean squared error and low R-squared error. The model is performing poor judging by the results.

We also test the K-neighbours Classifiers and we received an accuracy score of 75%.

We save Linear regression in .pkl format so we can deploy and evaluate the models with new data.

### Figure 29: Saving Linear regression model in .pkl format

## 

## Deployment

We have a separate file loan\_test datasets, which was used for deploying the Linear regression model. These datasets were with lesser number of rows and this time we didn’t want to use the drapna method, but instead we imputed the missing values with means. We also used LabelEncoder and StandardScaller. We added a column Loan\_status and we removed column Loan\_ID. Since the number of the columns from the test data matches with the number of columns with the training dataset, we could deploy the model. It made predictions that some of the instances are 98% predicted accurately.

### Figure 30: Linear regression deployed model

## Challenges

We followed best practices for remote collaboration, using Google Meet to facilitate meetings once weekly. We also utilised Google Docs to share documents, allowing us all to collaborate on the document. We followed CRISP methodology for working on the project, which was easy to follow.

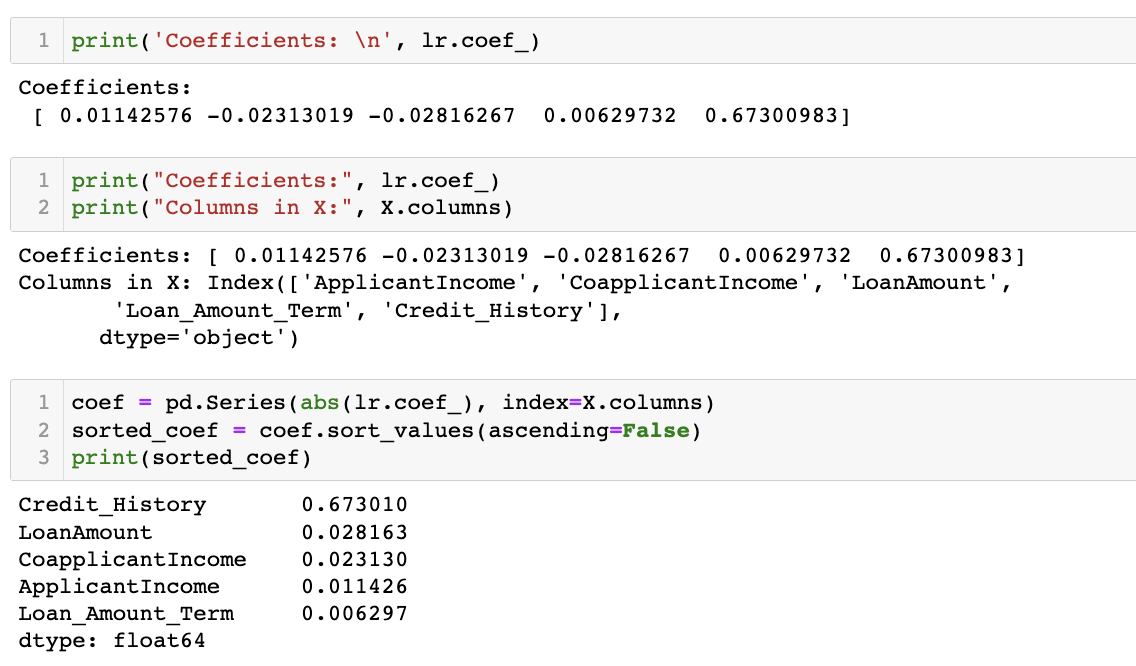
When we implemented the machine learning model, we received errors that there are still null values present. We repeated the process of cleaning the null values. The same applies for string data in column ‘Dependents’, where we treated the data with values 3+ and we needed to convert to integer. We had to repeat the steps.

It was challenging to implement the **deployment** of the machine learning model. We treated the new test data in a slightly different manner, where we imputed the missing values with ‘mean’ and also added a column ‘Loan\_Status’, as it was missing in the file. Even though we had all the libraries in place, we received an error that .pkl file is missing. After a couple of attempts, we managed to successfully deploy the Linear regression model and we received results that some of the instances are predicted with 98% accuracy.

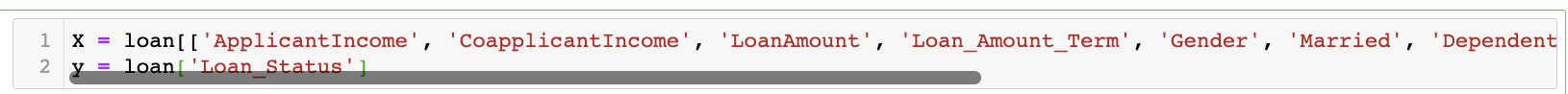
We outlined the feature importance for the best performing model and it has proven to be correct based on the research done by others in this area. We can not be fully confident that our data has been pre-processed with suitable methods or metrics, thus leading the outcome with biased results, however we used recommended approaches and accomplished satisfactory results.

## Results and analysis

Since the best performing score is for Linear regression model, we implemented coefficient magnitudes sorting values, based on the features we initially selected and we received the score as follows: Credit history 0.673, Loan amount 0.028, CoApplicant income 0.023, Applicant income 0.011 and Loan amount Term 0.006. We wanted to avoid biases and we added all of the attributes as features in this analysis and the outcome was slightly different.



### Figure 31: Coefficient magnitudes



### Figure 32: Additional features

The model was tested with all of the features present in the datasets Loan and ‘Credit history’ stands out with its feature importance.



### Figure 33: Ascending order of coefficients magnitudes

The highest ranking is Credit history, followed by the marriage status of the applicants and the request loan amount. The results from this dataset analysis confirm the findings of (Lee and Lee, 2018) that Credit history impacts the decision to lend loans.

It is important to highlight that Heatmap displayed the same results, where we noticed high correlation between Loan status and Credit history. This feature also influence on the requested loan amount, where in the event that the applicant doesn’t have credit history, there is a higher likelihood of rejection of the loan application and if the applicant has credit history, there is a higher likelihood of approval of the loan application.

We believe that these datasets are missing key applicant’s information. For instance, if the applicant is owning already a home or renting as the asset owning can affect the individual’s financial capacity.

Overall, we are content with the results and the outcome is sufficient and proves the hypothesis.

## Conclusion

Using a large dataset, this project has demonstrated how machine learning techniques can be applied effectively to predict loan defaults. After carefully inspecting and preparing the data, which included handling missing values, encoding categorical variables, and normalizing numerical features, we began our analysis.

We tested the predictive power of several machine learning models, including Linear Regression, Decision Trees, and Random Forests. We evaluated and compared these models using various metrics in a methodical manner. We were able to improve the performance of our linear regression classifier by using feature selection approaches, and it proved that credit history is the most significant predictor for approving loans to applicants. Other factors that influence the decision to lend loans include marital status. Individuals who are married are more likely to get their applications approved because, in case one of the individuals defaults on the loan, the partner can support and continue paying the due credit. Education also has an effect on the decision, and it appears to be the third most important feature.

## Reference list

Dabbura, I. (2019). *Predicting Loan Repayment*. [online] Medium. Available at: <https://towardsdatascience.com/predicting-loan-repayment-5df4e0023e92> [Accessed 6 April, 2023]

Harrison, M. (2019). *Machine learning pocket reference : working with structured data in Python*. North Sebastopol, Ca: O’reilly Media, Inc.

Hastie, T. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer*.

Lee, N. and Lee, V. (2018). *Bank Lending : Principles and Practice*. [online] Ebscohost.com. Available at: https://eds.p.ebscohost.com/eds/ebookviewer/ebook/ZTAyMG13d19fMTg3MDI2MV9fQU41?sid=471e9fd1-48c4-422b-b036-f17a767f9dfd%40redis&vid=1&format=EB&rid=1 [Accessed 5 Apr. 2023].

Konapure, R. (2023). *Home Loan Approval*. [online] www.kaggle.com. Available at: <https://www.kaggle.com/datasets/rishikeshkonapure/home-loan-approval>.[Accessed 29 March, 2023]

Müller, A.C. and Guido, S. (2017). *Introduction to machine learning with Python : a guide for data scientists*. Beijing: O’reilly.

Roy, B. (2020). *All about Feature Scaling*. [online] Medium. Available at: <https://towardsdatascience.com/all-about-feature-scaling-bcc0ad75cb35> [Accessed 9 Apr. 2023]