DEFAULT PREDICTION PROJECT REPORT

Background information

Business Understanding

**Problem Statement**

Create a credit scoring model to predict whether a customer will default or not

**Success Criteria**

To have a model with a high accuracy score

**Project Plan**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Plan** | **Duration** | **Resources** | **Input** | **Output** | **Dependencies** |
| **Problem definition** | 5 minutes | Project requirements | N/A | Definition of the project | Business understanding |
| **Success criteria** | 5 minutes | Project requirements | N/A | Success criteria of the project | Problem definition |
| **Project Plan** | 15 minutes | Crisp\_DM methodology | N/A | plan to achieve goals of the project | N/A |
| **Data sourcing** | 30 minutes | Project requirements and internet | N/A | Raw data acquired for the project | Problem definition |
| **Data preparation and Quality** | 2 hour | Python libraries | Raw Data | Data description report and quality data | Data sourcing |
| **Data cleaning** | 2 hour | Python libraries | Prepared data | Data cleaning report and final data to for analysis | Data preparation |
| **Data Analysis** | 1 hour | Python libraries | Clean data set | Analysis report | Data cleaning |
| **Modelling** | 5 hours | ML techniques | Training data | ML model | techniques |
| **Conclusion** | 15 minutes | Analyzed data | Analysis report | Conclusion to the project | Data analysis |
| **Recommendation and next step** | 15 minutes | Conclusion | Conclusions | Recommendation | Conclusion |

Data Understanding

**Data Sourcing**

The data was obtained from a challenge in an Afterwork Data Science and Moringa school meetup concerning credit scoring.

**Data Preparation and Quality**

The data was a csv file of 614 by 12.

 The data set summary was as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Data type** | **Missing values** | **Unique values** |
| Gender | object | 13 | Male, Female, nan |
| Married | object | 3 | No, yes, nan |
| Dependents | object | 15 | 0,1,2,3+, nan |
| Education | object | 0 | Graduate, not graduate |
| Self\_Employed | object | 32 | No, yes, nan |
| ApplicantIncome | int64 | 0 | “505  no.s” |
| CoapplicantIncome | float64 | 0 | “287 no’s” |
| LoanAmount | float64 | 22 | “203 no’s” |
| Loan\_Amount\_Term | float64 | 14 | 480,360,300,240,180,120,84,60,36,12,nan |
| Credit\_History | float64 | 50 | 1, 0 , nan |
| Property\_Area | object | 0 | Urban, rural, semiurban |
| Default\_Status | object | 0 | Y, N |

**Data Cleaning**

The missing values totaled up to 134 and hence it would be inappropriate to remove them. Other alternatives were sort, column wise:

For Loan Amount, the missing values were filled with 0. This was done also for Loan amount terms.The numeric variables co applicant income, loan amount term and loan amount were changed to integers.

The missing values in the credit history column were replaced with ‘nohistory’ while those in gender column were changed to ‘nonbinary’.

The entries with missing values in the married column were only three hence deleted.

The missing values in the self-employed column were filled with ‘temporary’ while those in the dependent column were filled with ‘unclear’.

There were no duplicate values.

Outliers were not removed.

**Data Analysis**

There seemed to be no outliers in the loan amount and loan amount term column and this was interpreted to mean that the term bracket is specified by the lending agency in question and so is the amount bracket; or the data is for these specific extremes.

Data Preparation

Constructed data

Modelling

Selected Techniques

Generated Test Design

Built Model

Parameter Setting

Model

Model Description

Model Assessment

Conclusion

Recommendation and Next Step