**Loan Default Prediction**

**Problem Statement:**

You are working within a finance firm as a data scientist. You are responsible for training and deploying the models that you create. As is often the case you are a part of the team that work towards the same goals. This creates a need to be able to collaborate with each other within the team and organization to quickly iterate. Training multiple models often leads to the need of managing and organizing the results, hyperparameters and storing the models. As you are a part of the team working in the same direction, it necessitates the need for transparency within the team to see which model everyone is trying, what works and what does not. To solve these requirements, the organization has decided to use MLFlow for all the data science teams. This will help them collaborate and iterate quickly, trying to solve a particular problem.

**Data Description:**

The dataset consists of 255347 observations and 18 features.

18 features are –

● LoanID - The loan ID

● Age - The age of the applicant

● Income - The income of the applicant

● LoanAmount - The total loan amount

● CreditScore - The credit score of the applicant

● MonthsEmployed - How many month has the applicant been employed

● NumCreditLines - How many credit lines

● InterestRate - The interest rate of the loan

● LoanTerm - The term of the loan

● DTIRatio - The debt-to-income ratio (DTI) ratio

● Education - The education level of the applicant

● EmploymentType - The employment type of the applicant

● MaritalStatus - The marital status of the applicant

● HasMortgage - If the applicant has mortgage or not

● HasDependents - if the applicant has dependents or not

● LoanPurpose - The purpose of the loan

● HasCoSigner - If cosigner exists

● Default - Did the applicant default

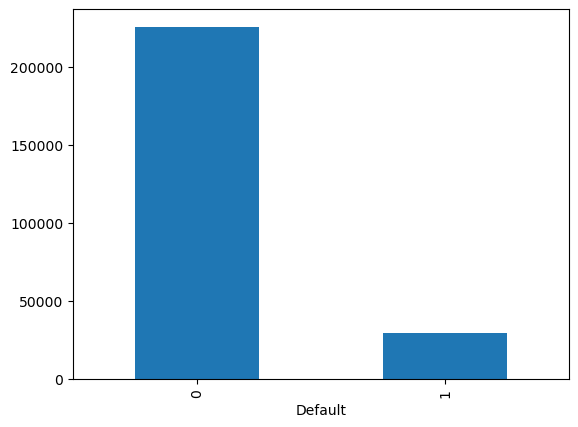
From the above 18 features “Default” is the Target Variable and rest all are Indeoendent Variables.

Data Types of Variables:

1. object – LoanID, Education, EmploymentType, MaritalStatus, HasMortgage, HasDependents, LoanPurpose, HasCoSigner
2. int64 – Age, Income, LoanAmount, CreditScore, MonthsEmployed, NumCreditLines, LoanTerm, Default
3. float64 – InterestRate, DTIRatio

There are no missing values in the dataset.

Understand Target Variable (Default):



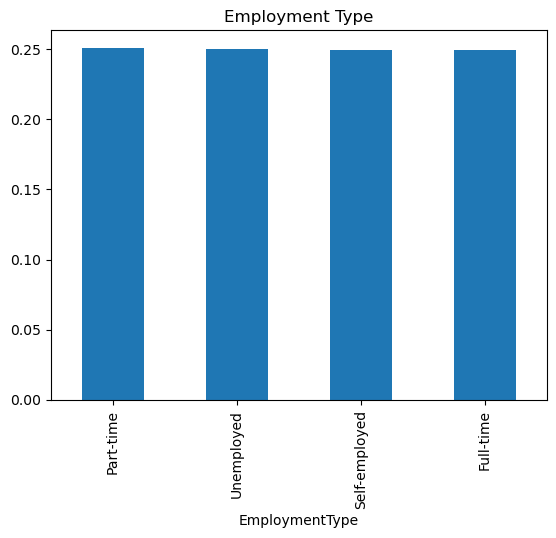
The target variable is highly imbalanced.

225694 (around 88%) loan applicants has not defaulted the loan out of 255347.

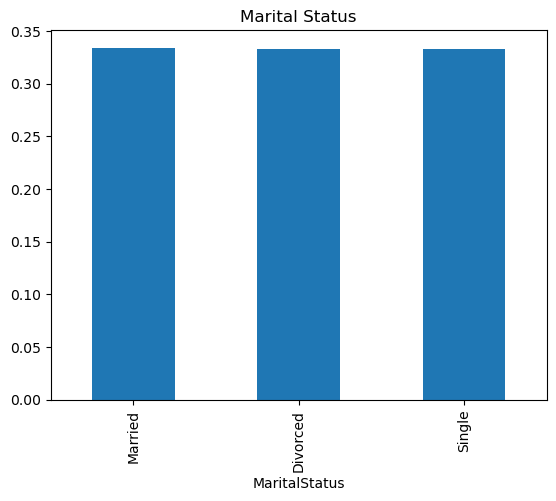
**EDA:**

**Univariate Analysis-**

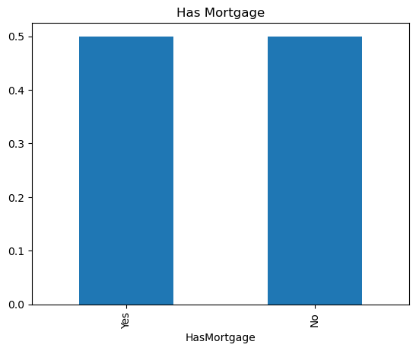
1. **Categorical Variables:**



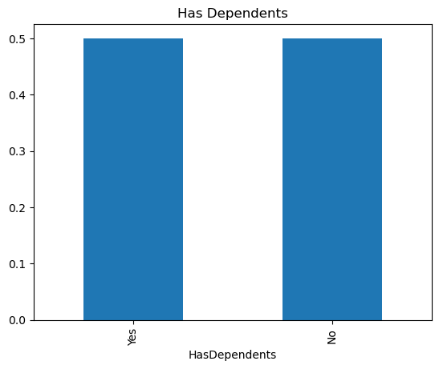
Count of Part-time, Unemployed, Self-employed and Full-time is almost same in dataset



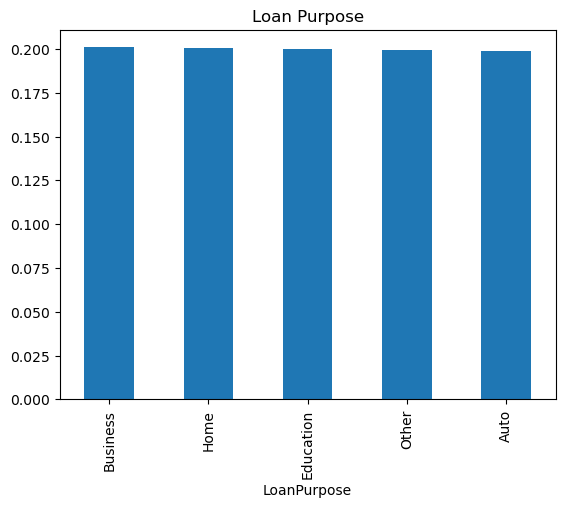
Count of Married, Divorced and Single are almost same in dataset



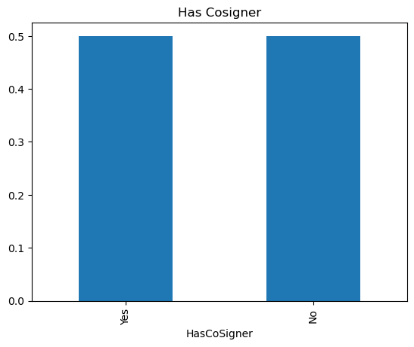
Count of Applicants with or without Mortgage is almost same



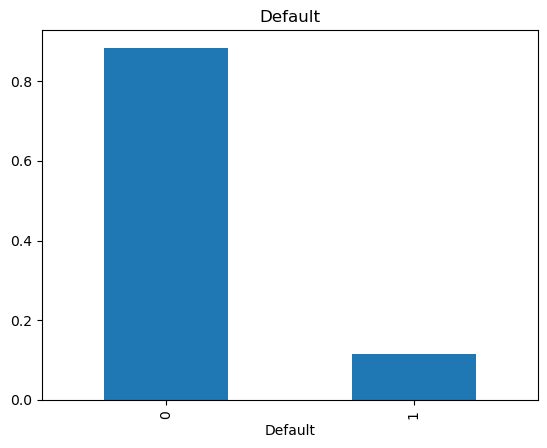
Count of Applicants with or without Dependents is almost same



Count of different loan purpose is almost same in dataset

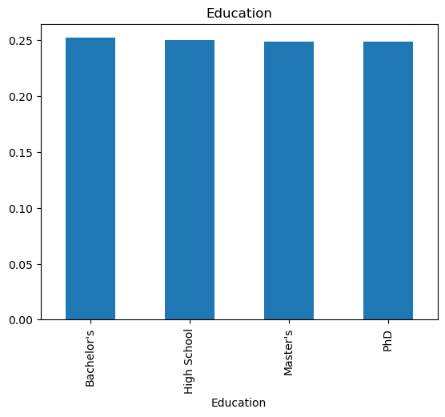


Count of Applicants with or without Cosigner is almost same

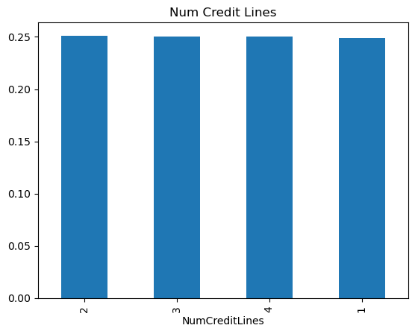


Less Loan Applicants has defaulted

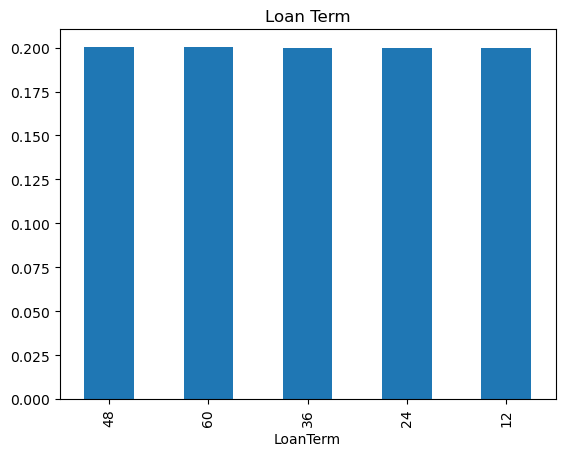
1. **Ordinal Columns:**



Count of applicants with different Education levels is same

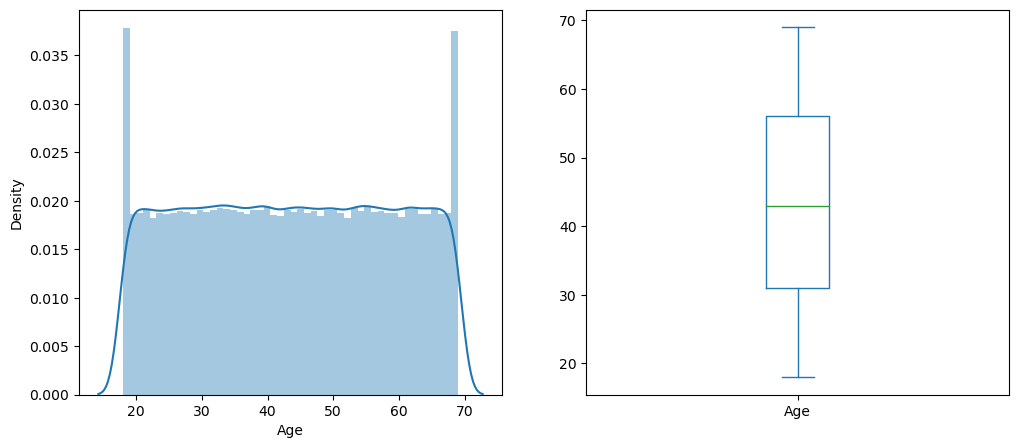


Count of applicants with different number of credit lines are almost same

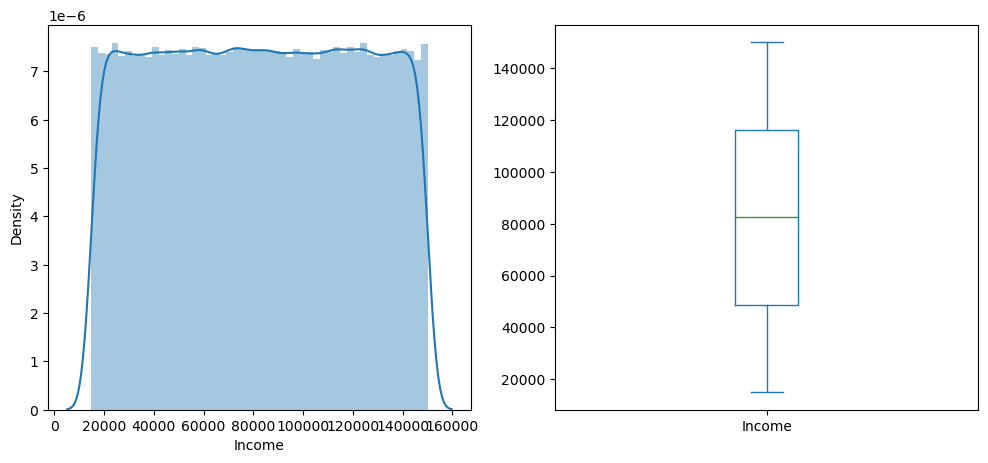


Count of Applicants with different years of loan term is almost same

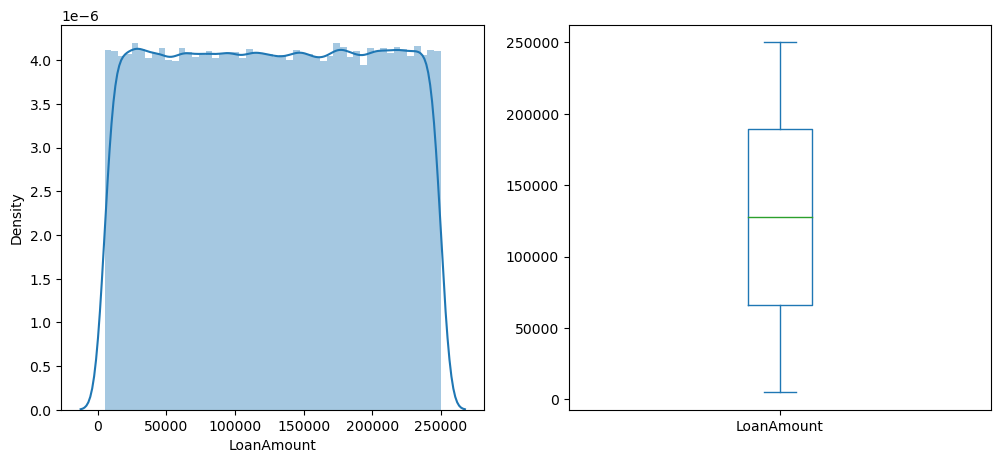
1. **Numerical Columns:**



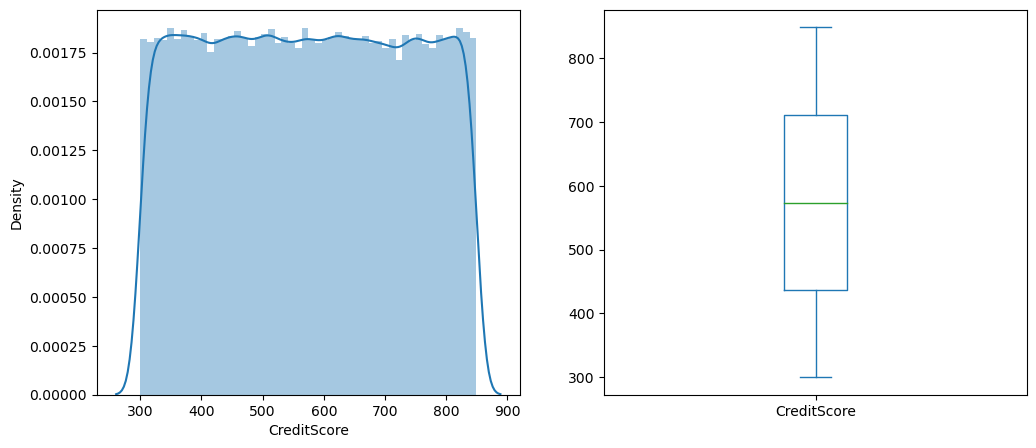
There is wide range of age among applicants in the dataset where most of them are in between 20 to 68



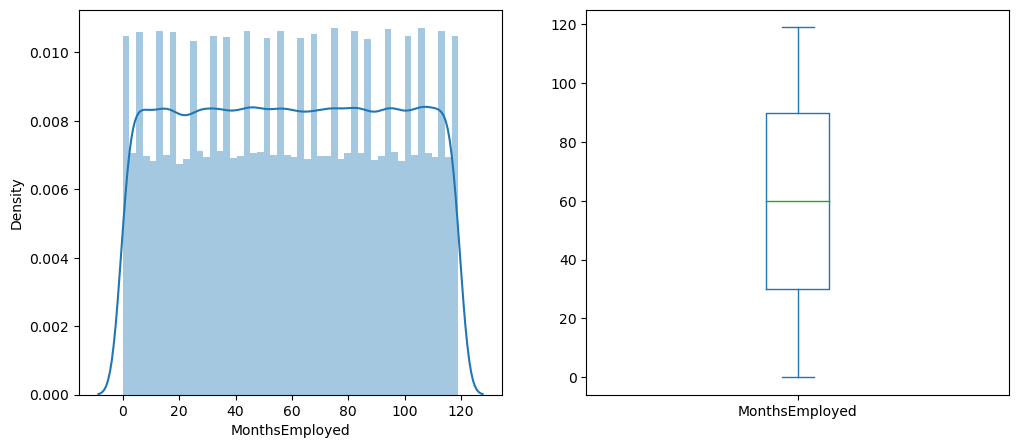
There is wide range of Income among applicants in the dataset with mean income around 80000



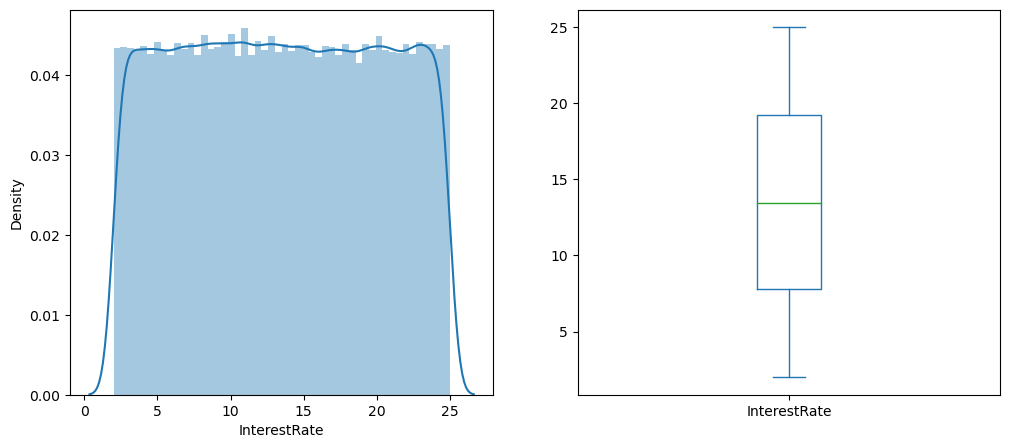
Mean Loan Amount is around 125000



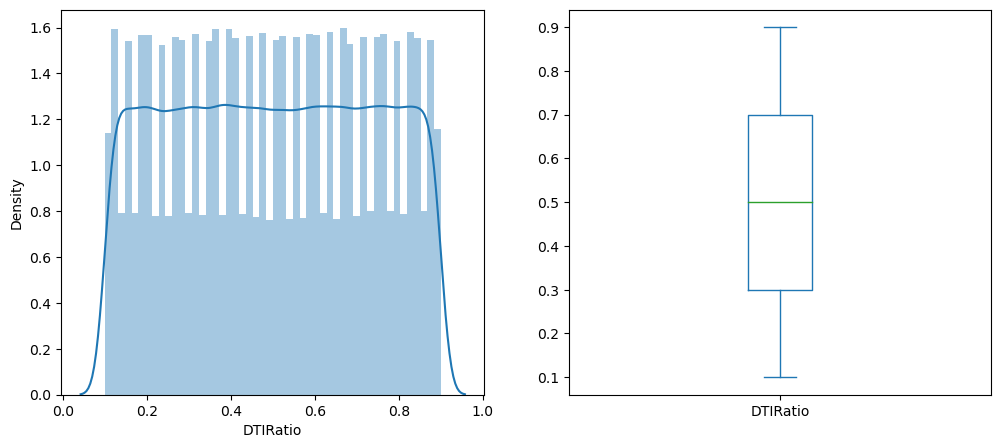
Applicants have wide range of Credit Score



Applicants have wide range of employment in months



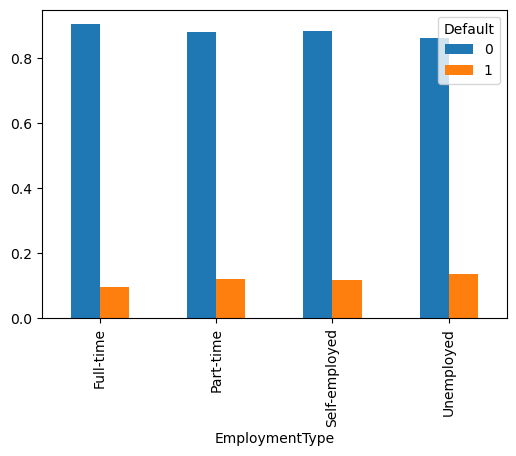
There are different interest rates based on loan of different applicants



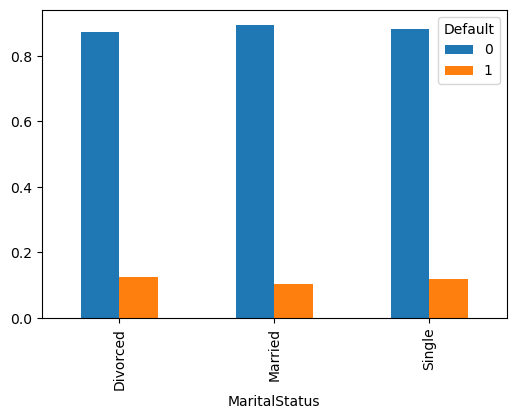
Different Debt to Income ratio based on applicants

**Bivariate Analysis-**

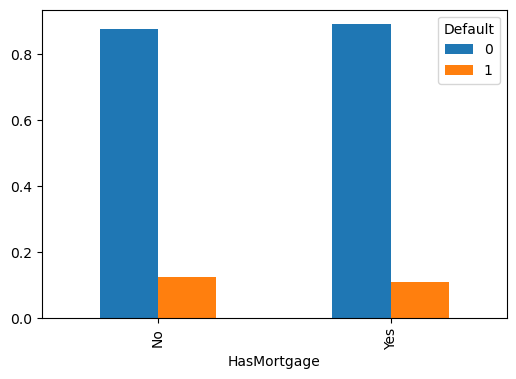
1. **Categorical Independent Variable vs Target Variable:**



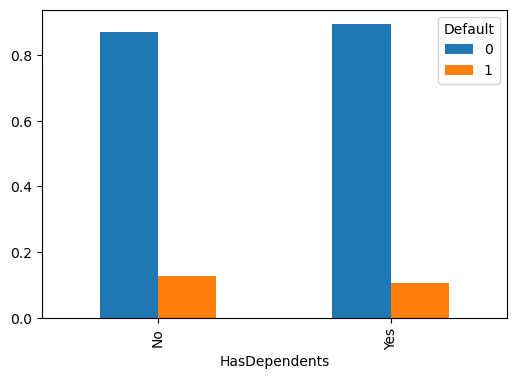
Unemployed are the most who defaulted the loan and Full-time employed are the most who has not defaulted the loan



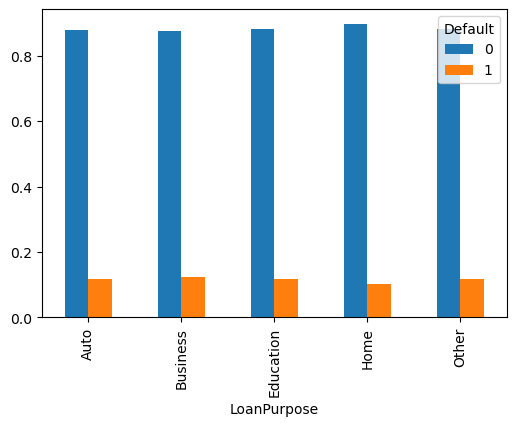
Married applicants are the most who has not defaulted as compare to Divorced and Single applicants



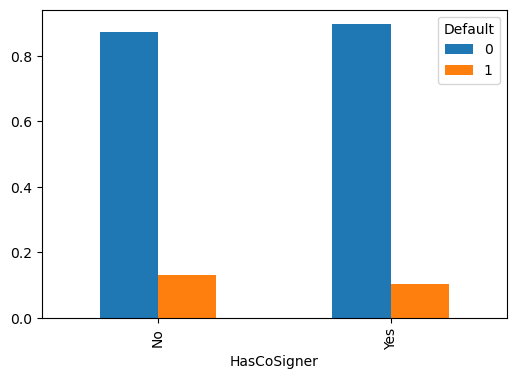
Applicants without mortgage has defaulted more as compare to those who have mortgage.



Applicants with no dependents has defaulted more.

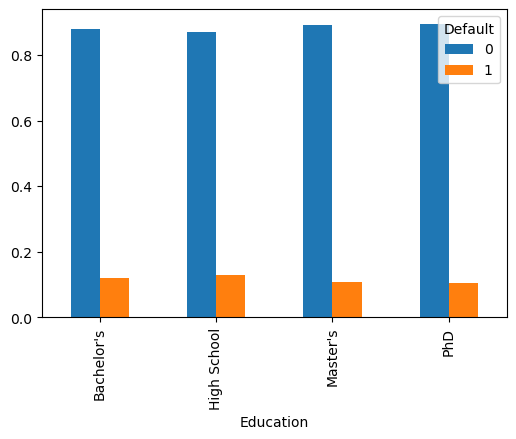


Applicants with Loan purpose as business, education and other has defaulted the most

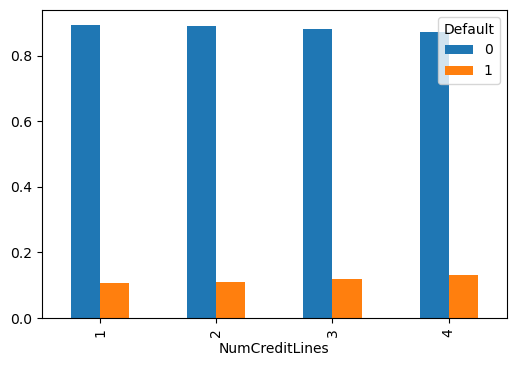


Applicants with no Cosigner has defaulted more

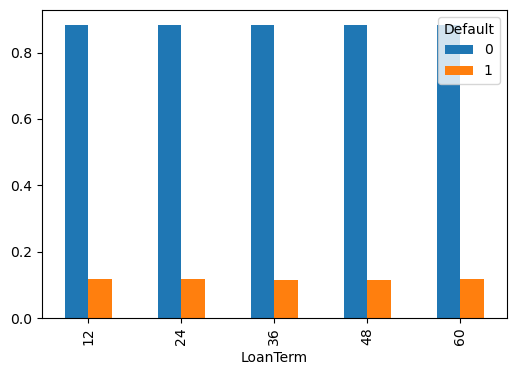
1. **Ordinal Independent Variable vs Target Variable**



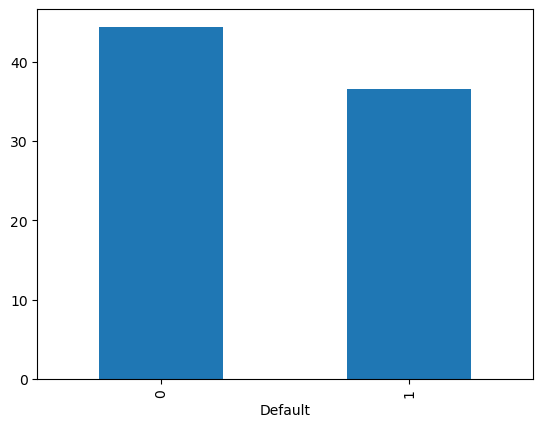
Applicants with High School degree has defaulted the most



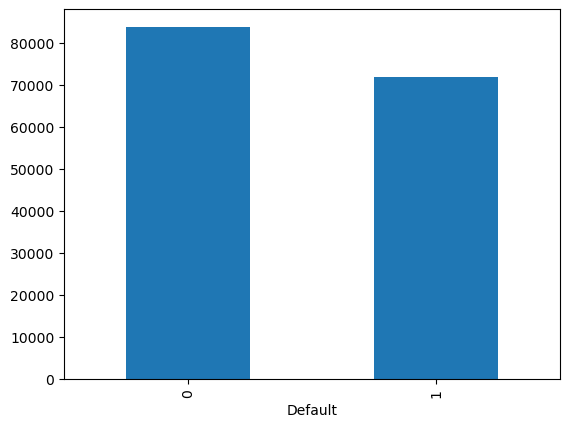
Applicants with Number of credit lines as 4 has defaulted more.



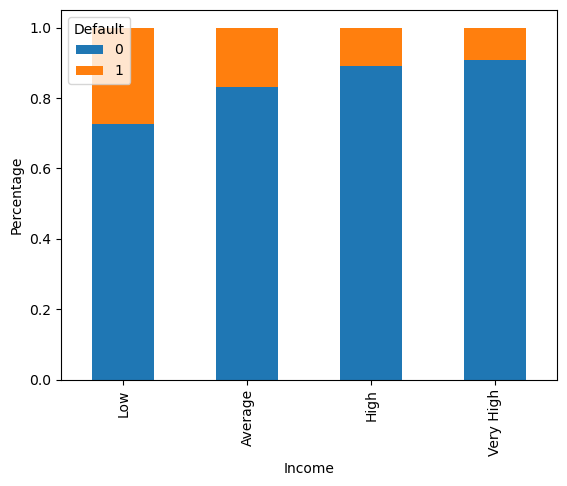
Loan Term does not specifically impact Default chance as Applicants with different loan terms has defaulted almost in same manner.



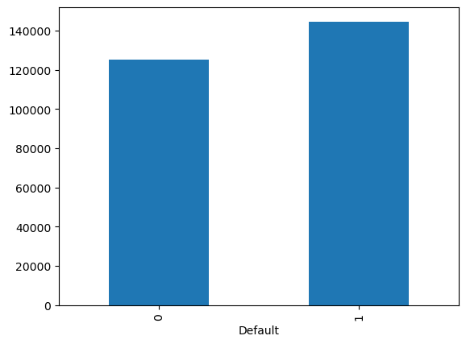
Mean Age of applicants who default is around 35 and the mean age of applicants who do not default is around 45



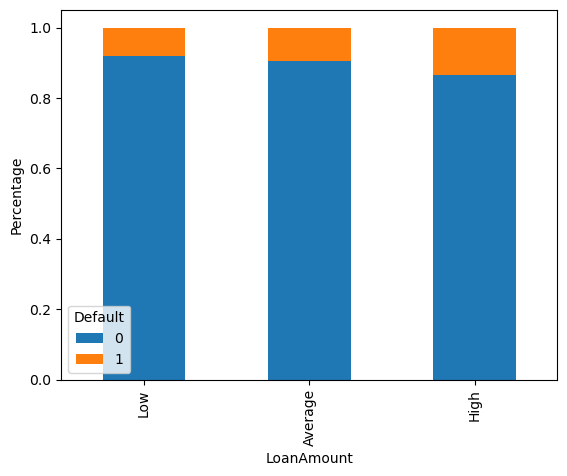
Mean Income of applicants who default is 70000



As the income of Applicants increases their default chance decreases

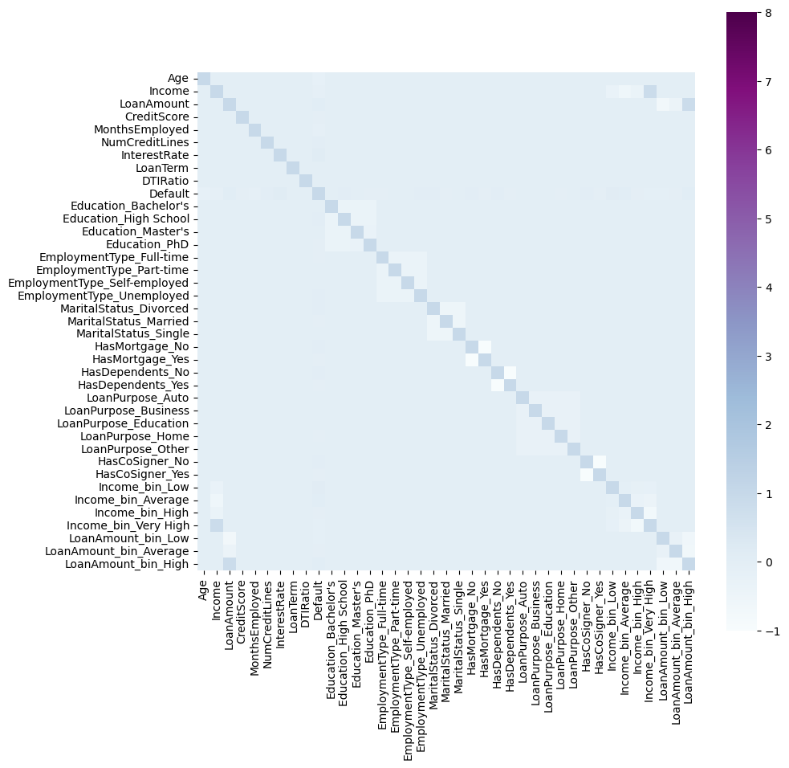


Mean Loan Amount where applicants default is around 120000



As the Loan Amount increses default chance also increases

Correlation Heatmap -



There is not high correlation among the variables so our dataset won’t face multicollinearity issue.

**Models Used:**

1. Logistic Regression

Parameters used : {"solver": "lbfgs", "max\_iter": 1000, "multi\_class": "auto", "random\_state": 80}

1. Bagged Decision Trees (Bagging)

Parameters used : {"estimator": None, "n\_estimators":8, "max\_samples":0.8, "max\_features":1.0, "n\_jobs":-1, "random\_state":80}

1. Random Forest Classifier

Parameters used : {"n\_estimators":10, "criterion":'entropy', "max\_depth":2, "min\_samples\_split":2, "min\_samples\_leaf":2, "min\_weight\_fraction\_leaf":0.0, "max\_features":'sqrt', "max\_leaf\_nodes":None, "n\_jobs":-1, "random\_state":80}

1. KNeighborsClassifier

Parameters used : {"n\_neighbors":3, "weights":'distance', "algorithm":'brute', "leaf\_size":30, "p":2, "metric":'minkowski', "metric\_params":None, "n\_jobs":-1}

1. Gaussian Naive Bayes

Parameters used : { "priors":None, "var\_smoothing":1e-09}

**Model Performance:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1 Score** |
| Logistic Regression | 0.8852 | 0.8512 | 0.8852 | 0.8356 |
| Bagged Decision Trees (Bagging) | 0.8803 | 0.8355 | 0.8803 | 0.8431 |
| Random Forest Classifier | 0.8844 | 0.7822 | 0.8844 | 0.8302 |
| KNeighborsClassifier | 0.8478 | 0.8053 | 0.8478 | 0.8239 |
| Gaussian Naive Bayes | 0.8848 | 0.8626 | 0.8848 | 0.8318 |

For our Loan default prediction project, False Negatives Rate or Recall is the best metric to evaluate the model. Lower the number of false negatives or higher the Recall Score, better the model is. In this project, False negative is when model predicting “a borrower will not default a loan even though he will “. Our model cannot afford having higher False Negatives as it leads to negative impact on the investors and the credibility of the company. So, we evaluated our models using the number of False negatives and accuracies.

**Logistic Regression is the best model for our project as it has highest Recall score.**

**MLFlow:**

Below is the screenshot of MLFlow where all the 5 models are saved along with their evaluation metrics.

