Loan repayment prediction

Using machile learning

Review -3

TABLE OF CONTENTS

Introduction

- With the increase in the banking sector, a mass of people is applying for bank loans but the bank due to its limited assets must grant to its assets to limited people only, so it would be a critical process for the bank to identify the safer options for granting loans.
- So, banking institutions wish to reduce the risk factor involved in the process. Our model is based on reducing the risk factor using machine learning.
- This model will help us identify the loan defaulters using the data mining algorithm

Data visualization and pre-processing

- This dataset is about past loans. The **Loan_train.csv** data set includes details of 346 customers whose loan are already paid off or defaulted.

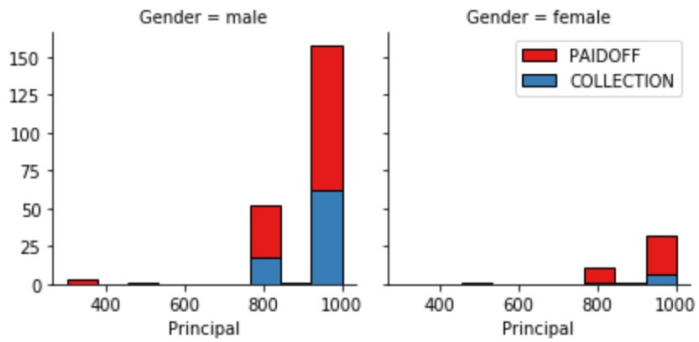
- Load Data From CSV File

- Convert to date time object

	Unnamed: 0	Unnamed: 0.1	loan_status	Principal	terms	effective_date	due_date	age	education	Gender
0	0	0	PAIDOFF	1000	30	2016-09-08	2016-10-07	45	High School or Below	male
1	2	2	PAIDOFF	1000	30	2016-09-08	2016-10-07	33	Bechalor	female
2	3	3	PAIDOFF	1000	15	2016-09-08	2016-09-22	27	college	male
3	4	4	PAIDOFF	1000	30	2016-09-09	2016-10-08	28	college	female
4	6	6	PAIDOFF	1000	30	2016-09-09	2016-10-08	29	college	male

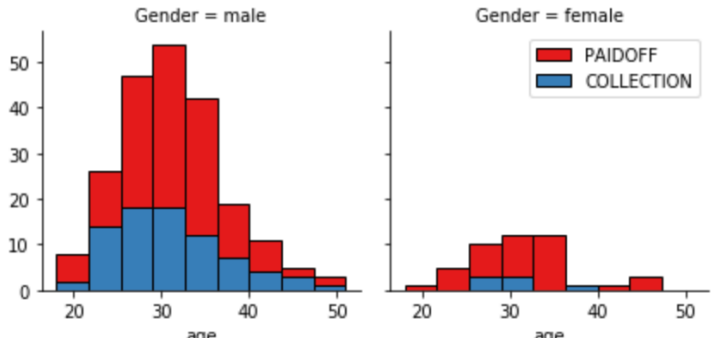
Data visualization and pre-processing

- Let’s see how many of each class is in our data set
- 260 people have paid off the loan on time while 86 have gone into collection
- Lets plot some columns to underestand data better:



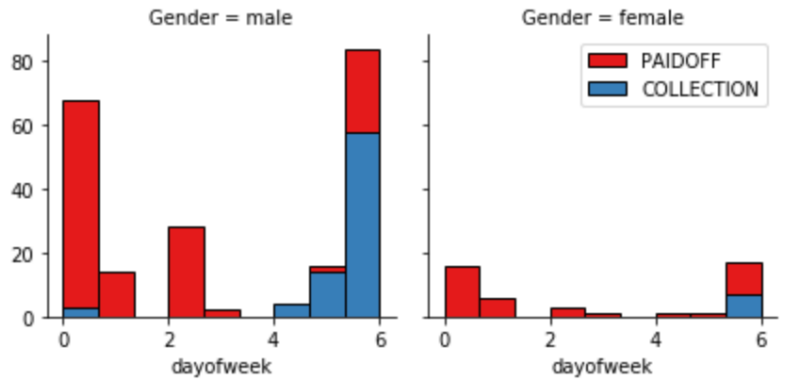
Pre processing

- Let us look at the age of the week people get the loan



Pre-processing

- Lets look at the day of the week people get the loan



Modelling techniques

- Logistic Regression :

Logistic regression is another supervised learning algorithm that is appropriate to conduct when the dependent variable binary. It is commonly used to obtain odds ratio in the presence of more than one explanatory variable. The procedure is quite similar to linear regression , but its response variable is binomial. The result is the impact of each variable on the odds ratio of the observed event of interest.

Support Vector Machine

- A Support Vector Machine (SVM) is also a supervised learning algorithm which used to separating hyperplane. In other words, given labelled training data, the algorithm outputs an optimal hyperplane which classifies new examples.
- In two-dimensional space this hyperplanes a line dividing a plane in two parts where in each class lay in either side

K-Nearest-Neighbours Model

- The k-nearest neighbours algorithm (KNN) is a non-parametric method that can be used for classification and regression problems. In classification problems, an object is classified by a vote of its neighbours, with the object being assigned to the class most common among its k nearest neighbours
- The prediction accuracy based on the k-NN model is highly contingent on the value of K. The best choice of K depends upon the data. Usually, larger values of K would reduce the effect of the noise on the classification but make boundaries between each category less distinct.

ALGORITHMS

- Import all the required python modules
- Import the database for both TESTING and TRAINING.
- Check any NULL VALUES are exists
- If NULL VALUES exists, fill the table with corresponding coding
- Exploratory Data Analysis for all ATTRIBUTES from the table
- Plot all graphs using MATPLOTLIB module
- Build the DECISION TREE MODEL for the coding.
- Send that output to CSV FILE.

Conclusion

- To conclude I would say that this model is considerably productive for the banking sector. Even being trained on small data has shown a positive response with great efficiency. If trained for real time data it can prove to be a real breakthrough in the banking sector.