Aim

Do exploratory data analysis on the data. Use different algorithms to predict the credit score. Identify the most important features in the data. Compare the findings from different methods.

Brief info about dataset

First, we have to import all the important modules and libraries in the notebook. The input data is of shape (100000, 28) and out of which we drop out 4 columns, namely, ID, Customer_ID, Name, and SSN because these features are not important for prediction. Your credit score doesn't depend on these 4 features. Also, there are no null values in any of the columns. Finally, we are left with 24 features and 100000 data points. Also, in the notebook, you can see the number of unique elements for each column.

Insights from dataset (EDA)

The number of standard credit scores remains approx. constant irrespective of the occupation with count around 3500. Also, the number of poor credit scores is highest for engineer and scientist occupations and the number of good credit scores is lowest for writer occupation. Thus, the majority of engineers and scientists have poor credit scores.

The feature credit mix is highly indicative towards the credit score. As one can see from the plots, with credit mix = good, the count of good credit scores is highest, with credit mix = Standard, the count of standard credit scores is highest and with credit mix = bad, the count of poor credit scores is highest.

Also, each type of loan has approximately the same count in the data with the minimum count of auto loans.

Skewness is a measurement of the distortion of symmetrical distribution or asymmetry in a data set. Most of the customer's are between 20 and 40 age and leading us to the skewness of 0.16. The less the skewness, the more is data close to normal distribution. Similarly, most of the monthly in hand salary is clustered towards 0-5000 units. Skewness of this column is much larger than the skewness of the age column. Also, we can conclude that most people who belong to the poor credit scores have mostly very small monthly incomes.

The number of bank accounts are closely related to the number of credit cards for a person, interest rate and number of loans a person holds. Also, most of the columns have zero covariance with the other columns. Annual income is highly correlated to the monthly income and amount invested monthly. Similarly, feature outstanding debt is correlated to interest rate and number of credit cards. All the features are real-life features and we can also think of correlation as similarity between different features like a person who owns more accounts will definitely have more credit cards.

If we talk about the number of credit score classes in the input data, 53.17 % are standard, 29 % poor and 17.83 % good. So, the dataset is somewhat skewed and biased.

The annual income column is highly indicative about credit score. For low income, most counts correspond to the standard and poor credit score while the high income corresponds to the good credit score. For intermediate incomes, mostly we have a standard credit score, and a tradeoff between good and poor credit score based on other features. The monthly in-hand salary follows the same as we can see from plots.

In the notebook, we can see density and box plot for each feature and based on the 1.5IQR rule, we can detect outliers and skewness of that feature.

Preprocessing the data

Now, we are doing preprocessing for the models. First for the categorical features we map them to the numbers, like, poor, standard, good credit scores are mapped to 1, 2, and 3 respectively. After this, we transform the input data or make mean zero and variance one for each data column and divide data into train and test datasets with 80:20 ratio respectively. Here, we have 24 input features and based on that, 3 output classes, namely, good, standard, and bad credit score.

K Nearest Neighbour (KNN) Algorithm

Then 1st we try KNN classifier model and train it. We set euclidean distance and vary the number of neighbors, or, k in it. Clearly, for k=1, we are achieving highest accuracy (training accuracy = 100% and testing accuracy = 76.14 %). For each class, we are getting precision, recall and f1-score in the range 70-79 %. From the confusion matrix, you can clearly see the number of wrong and right predictions.

Logistic Regression (LR) Algorithm

After that, we are trying a Logistic regression model which gives us the training accuracy = 64.14% and testing accuracy = 64.05 %. For each class, we are getting precision, recall and f1-score in the range 56-75 %. From the confusion matrix, you can clearly see the number of wrong and right predictions.

Comparison between KNN and LR model

From this, we can say that logistic regression is performing more poorly than KNN algorithm. KNN algo is non-parametric and supports non-linear solution in nature whereas Logistic Regression (LR) is parametric and supports linear solution. That is why KNN is slower than LR, but accurate than LR model. The precision of the LR model tampers with collinearity and outliers. Also, It can not be extended to problems of non-linear classification. KNN is one of the simple ML models, yet, it is a lazy model for learning, with local approximation and that is why, it has more accuracy over many supervised ML models. That is why, LR and KNN models are basically, trade-off between time requirement, memory requirement and accuracy.

Feature Importances

Also, monthly in hand income is the most valuable feature in the dataset for class poor credit score and annual income is least valuable. This is also shown by the plots during EDA. For standard credit score class, credit mix is most valuable as we can see that standard credit score comprises about 50% of data and for standard credit score, most of the count is of standard credit mix. For a good credit score, annual income is the most valuable feature and credit mix is of least importance. This is because of the fact that high annual incomes consist mostly of good credit scores as there are more chances to repay a loan but the good credit mix feature comprises both good and standard credit scores in the comparable amounts. We can view importances of all the remaining features from the feature importance barplot.

****