Final ML Project

RAJHA PRIYA A.

DTM10 Batch

Assessment answers:

1. Sales prediction with multiple variables using Multiple regression to predict the values and 4 assumptions. The linear model and smf statistics both models were prepared for this assessment
   For 5000 cpi, 3 percent discounts, 20 offers – sales are 826645.34
   For 4000 cpi, 8 percent discounts, 19 offers – sales are 732680.36

2. Comments on the result:
   - Model accuracy is 70% which is not bad, it will predict the loan application moderately
   - The major things considered for loan approvals for this model are cibil score and the cards held by the customer
   - There is a 30% of error seen in the model shows that a loan has been given to customers who are not eligible
   - F1 score is also noted as 72% which is good for predicting the loan category but still the specificity is higher which hinders the best performance of the model in my perspective
   - ROC AUC of our model approaches towards 1. So, we can conclude that our model does a good job of predicting whether we can give the loan or not to the customer
   - Overall, the model does a good job of predicting the outcome and the specificity is an important metric, especially when considering the balance between false positives and false negatives. Reducing the specificity could be better to obtain high accuracy in the model.

3. A) Decision tree – accuracy of the built-in model is 76% which is good and the probability in decision making is based on the occupation and sex to predict the income. Decision tree nodes with probability range distribution are attached in the final.ipynb

   B) Random Forest – accuracy is 80% and F1 score is 87% good classification RF model seems to be performing well. Also, RF showed better accuracy than the DT as compared. The consistency of scores across different folds suggests that the model is stable and not heavily influenced by training and testing in the dataset.

   C) KNN classification – 82% accuracy was observed in the data. KNN is successfully predicting the correct outcome most of the time.

   d) K means - I have implemented the most popular unsupervised clustering technique called K-Means Clustering. I have applied the elbow method and found that k=2 (k is a number of clusters) can be considered a good number of clusters to cluster this data. We got a relatively high accuracy of 50% for the model. But this model does not give the best performance. This Model has high interia not good fit.

   E) SVM – 75% accuracy obtained in the linear kernel which is good model.