Harnessing Machine Learning in Finance: Applications in Algorithmic Trading, Risk Management, and Fraud Detection

September 18, 2024

SPEAKER: MR. RAJVARDHAN RAVAT
Senior R&D Engineer | IPSoft Digital Inc.

SUBMITTED BY: SHUCHIKA SHARMA
Roll No: 24901325

DATA COLLECTION

A Loan Dataset was used which had a huge number of records. Some of the variables in the dataset include: average daily balance, deposit, number of years doing the business, person's credit score, social media presence, total postings etc.

GOAL - The goal is to find the things a person has to do to get his/her loan application accepted by the bank.

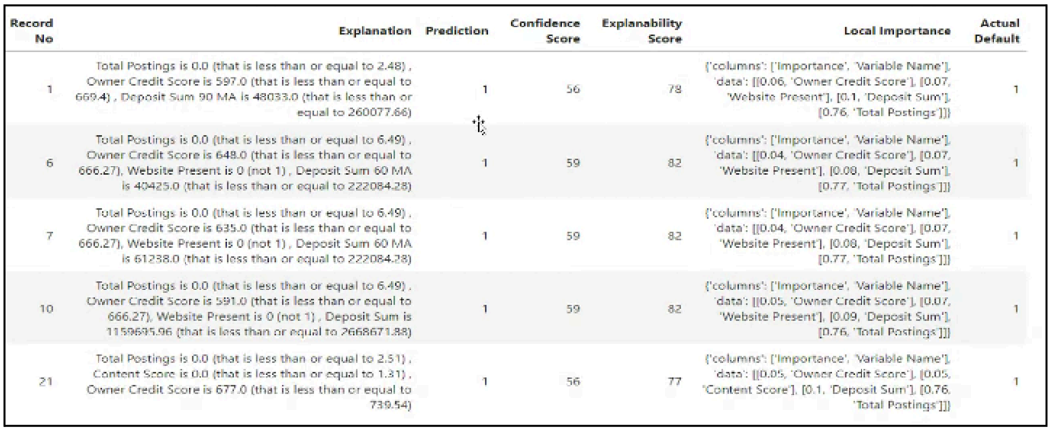
1. Data Quality Assessment - Below checks should be performed on the dataset before building any model -
    a. Shape - Dataset should be big enough to train the model
    b. Balance - Dataset should be balanced for each class in the classification dataset
    c. Emptiness - Check and remove missing values
    d. Outliers - Check for records for values which are out of the norm
    e. Completeness - Check if the data has the required information without any gaps
    f. Correctness - Check for human error or faulty devices/sensors
    g. Data Drift - Check for incorrect predictions due to data drift
    h. Influential features - Check for features which are more important
    i. Outcome Correlation - Check for co-dependent or redundant features and if they are strongly correlated to the outcome.
2. Alerting Setup On Dataset
    a. The bank would want to predict whether a person will be able to pay the loan before they give out the loan to the customer.
    b. For this purpose, the above data quality checks are first performed on the dataset.
3. Augmented Intelligence -
    a. In order to do more research on the dataset for model training, most influential features should be fetched from the dataset. These features will reflect those columns which affect the output column the most.
    b. Additionally, a score is assigned to those insights in order to show our confidence on those insights.
    c. For example - In the case study discussed in the session, the speaker covered the "Feature Insights of non-defaulters". Below insight was assigned highest insight score of 0.8 and an accuracy of 0.9 -
        i. Popular score > 0.5
        ii. Deposit Count 60 > 9.5
        iii. Total Postings > 22.5
        iv. Deposit Sum > 226406.0
    d. Ultimately, the feature selected to be with highest importance value was Total Postings, followed by footprint, Credit Score, Deposit Count and so on.

MODEL BUILDING

The speaker discussed the next step, which was model building and emphasised the importance of trying different algorithms to try and get the best metrics. Proper parameter tuning needs to be done to get the best accuracy. The performance metrics should also be checked on the test dataset to select the best model. The test dataset can be chosen via different methods like test-train split, k-fold etc. One-hot encoding should be performed on the dataset where necessary.

EXPLANATION OF PREDICTIONS

In this section, the speaker discussed the reason behind defaulter/non-defaulter predictions. For example, for a person with 0 total postings, credit score 597 and deposit sum 90, the model prediction was 1 (i.e. defaulter). This prediction was correct as these values were less than the thresholds of 2.48, 669.4 and 260077.6 respectively, which aligns with the manual identification as well during the ‘Feature Insight’ stage.



Record No	Explanation	Prediction	Confidence Score	Explanability Score	Local Importance	Actual Default
1	Total Postings is 0.0 (that is less than or equal to 2.48) , Owner Credit Score is 597.0 (that is less than or equal to 669.4) , Deposit Sum 90 MA is 48033.0 (that is less than or equal to 260077.66)	1	56	78	{'columns': ['Importance', 'Variable Name'], 'data': [[0.06, 'Owner Credit Score'], [0.07, 'Website Present'], [0.1, 'Deposit Sum'], [0.76, 'Total Postings']]}	1
6	Total Postings is 0.0 (that is less than or equal to 6.49) , Owner Credit Score is 648.0 (that is less than or equal to 666.27), Website Present is 0 (not 1) , Deposit Sum 60 MA is 40425.0 (that is less than or equal to 222084.28)	1	59	82	{'columns': ['Importance', 'Variable Name'], 'data': [[0.04, 'Owner Credit Score'], [0.07, 'Website Present'], [0.08, 'Deposit Sum'], [0.77, 'Total Postings']]}	1
7	Total Postings is 0.0 (that is less than or equal to 6.49) , Owner Credit Score is 635.0 (that is less than or equal to 666.27), Website Present is 0 (not 1) , Deposit Sum 60 MA is 61238.0 (that is less than or equal to 222084.28)	1	59	82	{'columns': ['Importance', 'Variable Name'], 'data': [[0.04, 'Owner Credit Score'], [0.07, 'Website Present'], [0.08, 'Deposit Sum'], [0.77, 'Total Postings']]}	1
10	Total Postings is 0.0 (that is less than or equal to 6.49) , Owner Credit Score is 591.0 (that is less than or equal to 666.27), Website Present is 0 (not 1) , Deposit Sum is 1159695.96 (that is less than or equal to 2668671.88)	1	59	82	{'columns': ['Importance', 'Variable Name'], 'data': [[0.05, 'Owner Credit Score'], [0.07, 'Website Present'], [0.09, 'Deposit Sum'], [0.76, 'Total Postings']]}	1
21	Total Postings is 0.0 (that is less than or equal to 2.51) , Content Score is 0.0 (that is less than or equal to 1.31) , Owner Credit Score is 677.0 (that is less than or equal to 739.54)	1	56	77	{'columns': ['Importance', 'Variable Name'], 'data': [[0.05, 'Owner Credit Score'], [0.05, 'Content Score'], [0.1, 'Deposit Sum'], [0.76, 'Total Postings']]}	1

SUMMARY

Below covers a brief summary of the session -

1. Data Quality checks are very important before performing the modelling.
2. In order to have a better understanding of the dataset before model building, feature importance and its impact on the results should be known.
3. Validation and test metrics are important in the model building step.
4. One should be able to explain the prediction given by their model for the model to be believable.