CASSANDRA

Team GREEKGODS

Private Score : 23.73101 (5th position)
Public Score : 24.40498 (13th position)

FEATURE ENGINEERING

1. Converted ['Created', 'Invoice_Date', 'Due_Date'] to datetime format
2. Filling Missing Values in 'Description':
   After a thorough analysis of the 'Description' column and using different techniques to fill the missing value, we sorted the data using the 'Created' column and then used the forward fill technique to fill the missing. This idea also correlates with the fact that while entering the invoice details on the accounting system, the 'Description' for continuous invoices having the same description doesn't need to be repeated.

3. Months and Years:
   We extracted the Months and Year from ['Created', 'Invoice_Date', 'Due_Date']. This idea is based on the fact that a specific month and year can tell a lot about a person's income and wealth, especially for businesses in which certain months are more profitable than others. This can greatly influence the time taken to make the payment.

4. Time Given To make the payment
   This feature will be a key feature in determinig the number of days till payment.

5. Encoding ['Description' ,'Vendor_Name'] columns using Target Encoding
   After reading, research, and experimentation, the best encoding technique we could find for encoding ['Description' , 'Vendor_Name'] columns is Target Encoding. We tried Cross-Fold Target Encoding and Leave-one-out Target Encoding, resulting in many Nan values. Since Vendor Name is one of the essential features for predicting the Number of Days until Payment, as it tells about the Vendor's history, we could not afford to lose any data. We decided to go with Target Encoding. We tried a few different features to do encoding, but Target encoding gave the best result.

6. Encoding the testing data
   Encoding the testing data using the 'Description' and 'Vendor_Name' dictionaries made using the training data. Since there were a few values in the testing data's Description, which were not there in the training data, we encoded such values using the mode of the training data's Encodings.

7. Additional Features

    Few of the additional features on which we thought the Number of Days until Payment will depend on are -

- 'Ven_num_di' : Average of Number of Days until Payment for each vendor (Target encoding of Vendor) / number of days given to the vendor to make the payment.
- 'Last_num' : Time taken by the vendor to make Payment on their Last Invoice sorted using 'Invoice_Date'. This feature will indicate the current financial condition of the vendor, which will further affect the number of days until payment.
- 'Last_num_di' : 'Last_num' / number of days given to the vendor to make the payment

8. Removing Outliers

    We intentionally did not remove any outliers from our training dataset. We did not want to lose any vital information about any vendor, and each data point added some information to vendor history. Moreover, the models we used in training were robust to outliers, and any experimentation with removing outliers only resulted in a lesser score.

9. Choosing Features for Training the models

    From the correlation heatmap and after lot of trial and error we dropped ['Due_Date','Description','Vendor_Name','Created','Invoice_Date','Last_num_di','Ven_num_di','Des_num'] features to avoid over-fitting the model.

Models and Training

- Multiple models were tried and implemented on the final training data including:
    - CatBoost,
    - XGBOOST,
    - RandomForestRegressor,
    - LGBM
- The Best scores were from an ensemble of CatBoost and RandomForestRegressor.
- Grid search was used to fine tune the CatBoost model
- We also tried addding a new feature to our training using the predictions from one model and passing them through another model but we were unable to get any significant imporvement in our score.