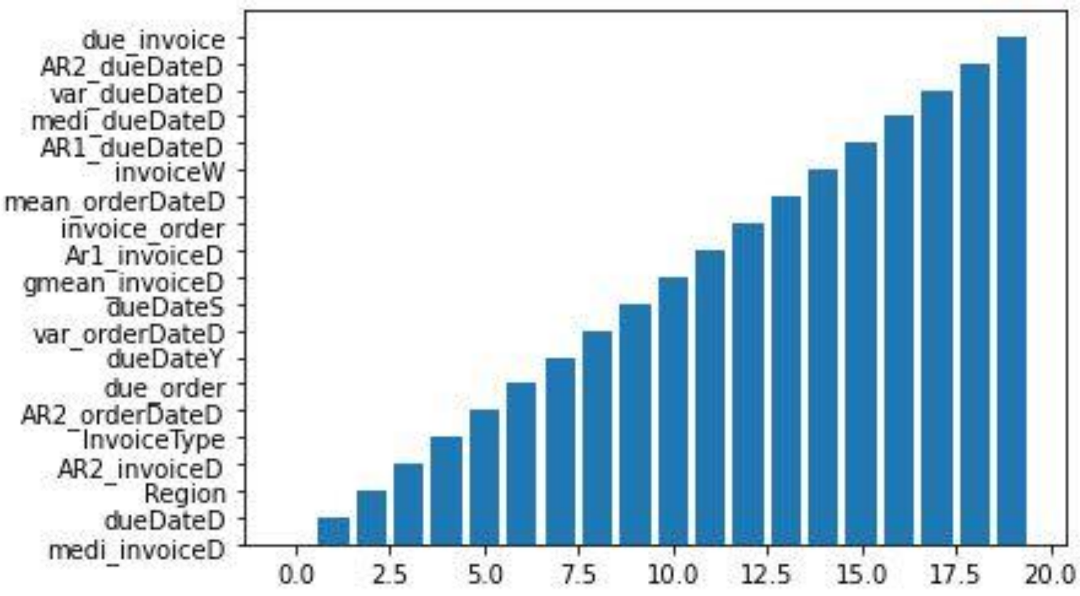
Overview:

A bank is analyzing the data of its customers and the input data is limited to 7 factors Invoice Type, due date, and geographical region. They are looking for a relationship between those factors and delays in the credit card bill payment. After preprocessing of the provided data from 1850 customers such as removing missing data, converting categorical data to numerical data, and finally extracting statistical information from the date data. Those 7 factors have been converted to 44 features or factors for analysis. Those data have been divided into two sets, train and test set, then the train set has been fetched to the machine learning model (XGBoost). The result of the feature ranking shows that (the below graph) the information due date and invoice date contain more informative information. This information can give more confidence to the bank that they can manage the customers who have delays in paying their balance late or help them to remind them.



Data Description

The dataset contains the following features:

- **UniqueId**: The unique identifier of an invoice.
- **Region**: The geographical region of the customer.
- **OrderDate**: The date on which the order was placed (YYYY-MM-DD).
- **InvoiceDate**: The date on which the invoice was generated (YYYY-MM-DD).
- **Amount**: The amount of the invoice (in US dollars).
- **DueDate**: The due date by which payment should be received (YYYY-MM-DD).

- **Dispute**: Indicates whether the customer has disputed the invoice amount (True/False).
- **InvoiceType**: The type of invoice ('electronic' or 'paper').
- **Delay**: The response variable indicating whether the payment was delayed (True/False).

Steps in the Notebook

1. Data Loading and Exploration

- Load the training dataset (train.csv).
- Display the first few rows and column names.
- Check for missing values and handle them by replacing with the mode of each column.

2. Feature Engineering

- Extract day, week, month, and season from the date columns (OrderDate, InvoiceDate, DueDate).
- Calculate the difference in days between DueDate and InvoiceDate, DueDate and OrderDate, and InvoiceDate and OrderDate.
- Generate various statistical features (median, mean, geometric mean, variance, standard deviation, covariance) for the date-related features.
- Drop the original date columns and UniqueId from the dataset.
- Convert categorical variables (Region, Disputed, InvoiceType) to numerical using one-hot encoding.

3. Model Training

- Split the data into training and test sets.
- Standardize the features using Z-score normalization.
- Train an XGBoost classifier using 10-fold cross-validation and evaluate the model using the F1-score.
- Fit the model on the entire training set.

4. Feature Importance

- Visualize the top 20 features based on their importance in the model.

5. Model Evaluation

- Load the test dataset (test.csv) and preprocess it similarly to the training data.
- Predict the Delay variable for the test set.
- Evaluate the model's performance using confusion matrix metrics (True Negatives, False Positives, False Negatives, True Positives).

- Calculate accuracy, precision, recall, and specificity.

Conclusion

This notebook provides a comprehensive approach to building a machine-learning model for predicting invoice payment delays. By following the steps outlined, users can preprocess their data, train an optimized model, and evaluate its performance effectively.