**INTERNAL OPS VOLUME FORECASTING DOCUMENT**

1. **DATA**

There are 2 data files stored in Excel format for the years 2022 and 2023:

+ Each data file will consist of 15 columns: 3 columns representing service types and 12 columns corresponding to the 12 months of the year for each service type.

+ There are only 2 data files, each file contains 12 records and 94 service types.

The data is too scarce to train a model to predict for the next 12 months of 2024. However, we are obliged to do so.

Requirement: From the data with 24 records, predict for the 12 months in the year 2024.

1. **FORECASTING FLOW**

+ Step 1: Data Processing:

- Load and handle missing or erroneous data.

- Create a column to concatenate the values of the columns "Business Service", "Delivery Service", and "LOB" together, separated by "=".

- Create an additional column to represent those services.

- Convert the data into a time series format with the index being the months of the years 2022 and 2023.

- Merge the data from the 2022 and 2023 files to obtain the following result:

+ Step 2: Calculate the correlation between columns in the dataframe: Based on the correlation matrix, we will identify which columns have strong correlations with each other (corr >= 0.08). Output the result to a text file.

+ Step 3: Build the prediction model: Before building the prediction model, it is necessary to check for stationarity and select an appropriate model.

+ Step 4: Convert the prediction data back to the original Excel format and export it to an Excel file.

1. **APPLIED ALGORITHMS**
2. **Arima**

+ Advantages: Suitable for models that do not require training and are non-seasonal. Easy to interpret. Provides a confidence interval for the prediction data.

+ Disadvantages: Requires a large amount of data. When the data is too scarce (such as 24 records in this case), it may lead to negative prediction results, which are incorrect.

1. **VAR**

**+ Advantages:**

+ Advantages:

- Developing from the ARIMA model allows for capturing the temporal dependencies and patterns in multiple variables simultaneously.

- Similar to ARIMA, this approach is interpretable and provides confidence intervals for predictions.

+ Disadvantages:

- Like ARIMA, it requires a sufficient amount of data to train the model effectively. Limited data (such as only 24 records in this case) can lead to unreliable predictions.

- Complexity increases when dealing with multiple variables, as the model needs to estimate parameters for each variable and their interactions.

- Interpretation of results may become more challenging due to the increased complexity of the model and the potential interactions between variables.

**3. Simple Exponential Smoothing (SES)**

+ Advantages: Suitable for data without training and without seasonal patterns. Provides prediction intervals for the predicted values. Cost-effective as it only stores the nearest data points. Still interpretable.

+ Disadvantages: More complex than ARIMA, as it requires optimizing the "smoothing constant - alpha" parameter for the model. Requires more data to be more accurate.

**4. Holt’s Exponential Smoothing/ Double Exponential Smoothing**

+ Advantages: Suitable for data with training but without seasonal patterns. Performs well with non-training data as well.

+ Disadvantages: Requires more data to be more accurate. Needs to optimize the alpha parameter for better model performance - typically choose alpha = 0.2.

**5. Holt’s Winter Exponential Smoothing/ Triple Exponential Smoothing**

+ Advantages: Suitable for data with training and seasonal patterns. Performs well with non-training data as well.

+ Disadvantages: Requires more data to be more accurate. Needs to optimize the alpha parameter for better model performance - typically choose alpha = 0.2.

1. **RESULT**

In this problem, the first step is to check for stationarity: Most of the columns are stationary => choose the Holt’s Smoothing method to build the model.

The prediction results for most columns are expected to be quite close to the manually predicted results in the existing Excel file, and some columns may contain negative values (columns with negative values are not suitable for this model):

We will separate the prediction columns into two dataframes: one for columns predicted to have negative values and one for columns predicted to have non-negative values.

For the dataframe containing columns predicted to have negative values, we will use Holt’s Winter method for prediction. This is because these columns may contain seasonal patterns.

The prediction results obtained for all 12 months of 2024 are stable around a certain value:

In summary, after analysis, we have utilized two prediction methods: Holt and Holt’s Winter for this dataset. The obtained results will be combined with other models such as Random Forest,... to improve the prediction accuracy. This helps enhance the diversity and reliability of the final predictions.