LTFS 2 Hackathon

Approach

- Simple and easy approach applied to solve this hackathon
- Handled both segments separately and applied different models to forecast for each segment
- Data preprocessing was applied differently to each segment after understanding the seasonality and patterns in the data
- Aggregated at date level and added case counts for segment 1 and applied Holt's winter technique to get the forecasts
- Cleaned data for segment 2 to remove noise and aggregated at date level to add case_counts and generated date level features such as day, week, month, year, etc and target encoding column based on year.
- Holts winter technique worked for segment one and gave lowest validation MAPE
- Ensemble modeling with ExtraTreesRegressor and HistGradientBoostingRegressor as base models with default parameters and using LinearRegresssion as final model worked for segment 2 forecasting and and gave lowest validation MAPE.

Data-preprocessing / Feature engineering

Segment 1

- Filtered data for segment 1
- No preprocessing for segment 1 data
- Transformed the data by aggregating at date level (application_date) and added the case counts across branches by day

Segment 2

- Filtered data for segment 2
- Removed data related to Punjab and Haryana as they have zero case counts to reduce noise in the data
- Noticed the records for segment 2 are made consistent to start on the same start date and imputed the data with zeros
- Cleaned data set to remove records for each state with early zero case counts in the data
- Finally, aggregated at date level (application_date) and added the case counts across branches by day.
- Generated date based features such as day, week, month, year, etc and did target encoding on year column and ended up with 17 features in the train set.

Final Model

After struggling almost six days with trying numerous complex methodologies for forecasting the data, I finally started applying simple methods that worked for solving this hackathon and getting good place on the leaderboard.

- Applied Holt's winter method because of the seasonality factor on segment one and achieved the lowest MAPE 12.866 on my Validation set compared to other models.
- Ensemble modeling with ExtraTreesRegressor and HistGradientBoostingRegressor as base models with default parameters and using LinearRegresssion as final model worked for segment 2 forecasting and gave me the lowest MAPE 13.973 ON Validation set compared to other models.
- Simple modeling strategy gave me best results (20.373 on PVT leaderboard).

Key Takeaways

- Good chance to learn and implement forecasting techniques on the real world data
- Data needs to be prepared separately for each segment based on the trends and seasonality
- Feature engineering needs to done based on the data segment and date based features such as year, week, day, month, day of week, etc works in order to capture trends
- Single forecasting model won’t work in all cases and have to be applied by understanding the data and patterns
- Holts winter
- Ensemble modeling with ExtraTreesRegressor and HistGradientBoostingRegressor as base models and using LinearRegresssion as meta model worked for segment 2 forecasting.
-

External Datasets

- I didn’t had a chance to use any external data sets for this dataset.