**Approach Note**

1. Problem Statement: To forecast the demand in the car sales so as to keep the inventory ready for unexpected demands
2. Types of solutions:
   1. Time series forecasting methods:
      1. Holts Winters
      2. ARIMA
      3. SARIMA
      4. SARIMAX
      5. LSTM
      6. FB Prophet etc.
   2. Regression models:
      1. Linear Regression
      2. Decision Tree Regression
      3. Random Forest Regression
      4. XGBoost
      5. LightGBM
      6. **CatBoost** (used this approach to solve this problem)
3. Following are the steps taken to approach this problem:
   1. Import necessary packages
   2. Read data
   3. Exploratory data analysis (Data types, structure, missing values, Correlation etc.)
   4. Train-validation-test split
   5. Catboost regression
   6. Predict and evaluate
   7. Hyper parameter tuning
   8. Feature enhancements (holiday list, addition of quarter, daysofyears etc.)
   9. Submission
4. Following are few main takeaways from this assignment
   1. Used pandas data manipulation technique to **create datetime column** from date and hour column
   2. Once the data has been imported, exploration of data is a necessary step and the important insight we found out in this data is that there were **missing values from 27th November 2020 and 27th December 2020**. Assuming that it was COVID strict lockdown.
   3. Since there was this missing piece, I had to divide the train-validation-test data into following:
      1. **Test Set** – All data after 27th December 2020
      2. **Train Set** – All data between 18th August 2018 and 31st July 2020
      3. **Validation Set** – All data between 1st August 2020 to 27th November 2020
   4. Since my approach was to create regression model using catboost, I had to disintegrate the time field into multiple features:
      1. 'datetime\_day',
      2. 'datetime\_week',
      3. 'datetime\_month',
      4. 'datetime\_quarter',
      5. 'datetime\_year',
      6. 'datetime\_hour',
      7. 'datetime\_minute',
      8. 'datetime\_second',
      9. 'datetime\_weekday',
      10. 'datetime\_dayofyear',
      11. 'datetime\_day\_name',
      12. 'datetime\_is\_month\_start',
      13. 'datetime\_is\_month\_end',
      14. 'datetime\_is\_quarter\_start',
      15. 'datetime\_is\_quarter\_end',
      16. 'datetime\_is\_year\_start',
      17. 'datetime\_is\_year\_end',
      18. 'datetime\_is\_leap\_year',
      19. 'datetime\_days\_in\_month',
      20. 'is\_holiday'
   5. Checked for correlation between these features and target.
      1. High correlations with target were first trained and tested with
      2. Then eventually removed the ones with high collinearity
      3. And finally selected 4 important features:
         1. **Hour**
         2. **Day**
         3. **Weekday**
         4. **Day of Year**
      4. **Holiday feature** did not have much feature importance from catboost model as well as did not have much correlation with the target.
   6. Created a Catboost Regression Model, with a lot of hyper parameter tuning.
   7. Predicted the test set, and evaluated using evaluation metrics like **RMSE, MSLE, MAE & R2** (for goodness of fit)
   8. Submission dataframe was created and then the csv was stored as per the require schema
   9. Manually kept uploading these csv on to the portal to get the score
   10. Used a lot of permutation combinations to reduce the test RMSE on the portal:
       1. **Changed model’s hyper parameters**
       2. **Feature Enhancements**
       3. **Used different slicing and dicing**

**Please Note:**  I have attached a solution.csv in this zip folder. It has a RMSE of 33.65 on the private test data set. I had uploaded exactly at 25th April 00:00, hence it is not visible on the portal. Please check.  
The approach here was simple, I used Grid Search CV which is in the code file in the zip folder, took the best parameter and extracted the CSV file. I got this result of RMSE 33.65, which makes me 6th on the leaderboard. Please do consider.

Thank you.

Saurabh Nair