**Nested CV**

**Data Preprocessing:**

Test data has 7560 rows and 12 columns while Training data has 1080 rows and 12 columns. After that I have done **EDA** on the data and after that I have added date and time in it. I Have check the null values in it . I extracted features like year, month, day of week, and whether it's a weekend.

In the **Feature engineering** , I have encoded categorical variables into numerical using label encoding. Also created lag features, rolling mean, rolling variance, differences, seasonal differences, rate of change, cumulative sum, and exponentially weighted moving averages.

**Modeling:**

I have imported various regression models from scikit-learn, XGBoost, CatBoost, and LightGBM, etc. After that I performs cross-validation for each model on a simplified set of features (`X\_train\_simple`) using R² score as evaluation metric and got the result of some of them are as likely Linear Regression: 0.1308 , 0.2695 , Ridge Regression: 0.1310 , 0.2692 , XGBRegressor: 0.8100 ,0.1022 of the Cross validation score and R² score respectively.

After applying CatBoostRegressor model and evaluating the CatBoostRegressor on the test set and prints out various regression metrics such as R² score, mean absolute error, mean squared error, root mean squared error, max error, and mean absolute percentage error. **Тrain: 0.9997599601157552**

**Тest: 0.9745586540639505 ,r2\_score: 0.9745586540639505 ,mean\_absolute\_error: 1325.9500221439614 ,mean\_squared\_error: 8151613.353773444, root\_mean\_squared\_error: 2855.1030373304293, max\_error: 45809.85269749881 ,mean\_absolute\_percentage\_error: 0.04455476217033474**

After analyzing that the plots feature importance for the CatBoost model and I can say that rolling\_mean\_3 has the highest feature importance.

Also , applying the model CatBoostRegressor on the X\_test\_shop\_1. And got the result, **r2\_score: 0.9317967966488715,mean\_absolute\_error: 2114.561870876629, mean\_squared\_error: 18045192.84504017, root\_mean\_squared\_error: 4247.96337614158 ,max\_error: 27217.78822952818, mean\_absolute\_percentage\_error: 0.0397277554787647.**

**Visualization:**

I have made the plots the actual vs. predicted quantities over time for the entire test set upto 2018. Also plots the actual vs. predicted quantities over time for specific shops in a specific city.

Again , applying the CB model on the X\_test\_shop\_2. And got the result of r2\_score: 0.9902460052547786,mean\_absolute\_error: 1066.3060806773728, mean\_squared\_error: 2299221.4516646564, root\_mean\_squared\_error: 1516.3183873001924, max\_error: 4792.019515904663,mean\_absolute\_percentage\_error: 0.030708630241999917

**Analysis and Interpretation:**

After evaluating model performance using various metrics and visualizations. Provides insights into the forecasted quantities for specific shops in a particular city. I got the forecasting result .

