AMAZON ML HACKATHON

What I did for featuring engineering?

In this code I am performing, feature engineering on a dataset by extracting size information from the 'TITLE' column using regular expressions and creating new columns based on the extracted data. The steps involved in my pre-processing approach was:

1. Reading the dataset from a CSV file using pandas library and creating a copy of the dataframe.

2. Combining the 'TITLE', 'BULLET\_POINTS', and 'DESCRIPTION' columns to create a new 'TITLE' column in the dataframe, and dropping the original columns.

3. Calculating the number of empty strings in the 'TITLE' column and grouping the data by 'PRODUCT\_TYPE\_ID' to identify the top 1% of products based on their frequency.

4. Filtering the original dataframe to keep only the rows with 'PRODUCT\_TYPE\_ID' in the top 1% list.

5. Defining a regular expression to extract size information from the 'TITLE' column and creating a WordNetLemmatizer object for lemmatization.

6. Extracting size information from the 'TITLE' column using the regular expression and storing it in a new 'SIZE\_INFO' column.

7. Lemmatizing the 'SIZE\_INFO' column and storing it in a new 'LEM\_SIZE\_INFO' column.

8. Removing the 'SIZE\_INFO' column from the dataframe.

9. Storing the resulting dataframe in a CSV file.

10. Reading the CSV file and creating a new 'SIZE\_NUM' column to store the numerical values extracted from the 'LEM\_SIZE\_INFO' column using a regular expression.

11. Removing the 'LEM\_SIZE\_INFO' column from the dataframe and storing the resulting dataframe in a CSV file.

12. Converting the 'PRODUCT\_TYPE\_ID' column to a categorical type, filling missing values in the 'SIZE\_NUM' column with 0, and removing rows with NaN values in the 'SIZE\_VEC' column.

13. Reading the CSV file and converting the 'SIZE\_NUM' column to a list of floats, sorting the list in descending order, and creating new columns with the top 3 values, filling with 0 if there are less than 3.

14. Dropping the original 'SIZE\_NUM' column from the dataframe.

15. Also LabelEncoding the PRODUCT\_ID\_TYPE and dividing it by len of classes

WHAT MODEL DID I USE?

My code is using the CatBoostRegressor model to build a machine learning model for regression. The model is being trained with the training data (X\_train and y\_train) using the specified hyperparameters: loss\_function is set to 'MAPE', random\_seed is set to 42, iterations is set to 1000, learning\_rate is set to 0.1, and depth is set to 6.

The model is being fit with the categorical features (cat\_features) and evaluated on the test set (X\_test and y\_test) using the MAPE (mean absolute percentage error) as the loss function. The early\_stopping\_rounds parameter is set to 50, which means that the model training will stop if the model's performance on the test set does not improve after 50 rounds of training.

Finally, the use\_best\_model parameter is set to True, which means that the best model based on the evaluation metric (MAPE) on the test set will be used.