**ML CHALLENGE SUBMISSION FILE**

APPROACH

The problem statement demands to predict the length of a product from catalog metadata, we need to develop a machine learning model for the same. This is a regression type problem, where the target variable is continuous. There are various regression algorithms like Linear Regression, Random Forest Regression, Gradient Boosting Regression, etc. To select the best algorithm, we can have used K-Fold Cross Validation technique. Using this method, the model will select the algorithm with the best performance on the validation set.

FEATURE ENGINEERING

The dataset provides various features like product title, description, bullet points, product type ID, and product length. We have used NLP (Natural Language Processing) techniques to extract useful features from the text data. Some of the features that we have extracted from the text data are:

1. Word Count: The number of words in the product title, description, and bullet points. This feature is used to further get the product length.

2. Word Length: The average length of words in the product title, description, and bullet points is used to get insights into the complexity of the product description, which can help in predicting the product length.

3. TF-IDF: Term Frequency-Inverse Document Frequency (TF-IDF) is a numerical statistic that reflects how important a word is to a document in a collection or a corpus. TF-IDF is implemented using a tokenizer by breaking down the text into tokens or words. These words are then converted into numerical vectors that are used in the model.

TOOLS

For the given problem statement, we have used Python programming language and various machine learning libraries like Scikit-learn, Keras, PyTorch, and Transformers (Hugging Face library) along with these for NLP tasks we are using the BERT language model. For deployment and testing, we have used Google Colab and GitHub for version control and collaboration.