**Overview**

The current pandemic has dwindled the data science job market likewise recruiters are also facing difficulties filtering the right talent. To bridge this gap we bring a chance for the MachineHack community to compete for jobs with some of the key analytics players for a rewarding career in Data Science. In this competition, we are challenging the MachineHack community to come up with an algorithm to predict the price of retail items belonging to different categories. Foretelling the Retail price can be a daunting task due to the huge datasets with a variety of attributes ranging from *Text, Numbers(floats, integers), and DateTime*. Also, outliers can be a big problem when dealing with unit prices.

With a key focus on the Data Scientist role in an esteemed organization, this hackathon can help freshers and experienced folks prove their mettle and land up in a rewarding career.

By participating in this hackathon, every participant will be eligible for the Data Scientist job role by making sure their MachineHack Information with Resume is up to date.

**Dataset Description:**

* **Train.csv - 284780 rows x 8 columns *(****Inlcudes****UnitPrice****Columns**as**Target****)***
* **Test.csv - 122049 rows x 7 columns**
* **Sample Submission.csv -**Please check the **Evaluation** section for more details on how to generate a valid submission

**Attribute Description:**

* **Invoice No - Invoice ID, encoded as Label**
* **StockCode - Unique code per stock, encoded as Label**
* **Description - The Description, encoded as Label**
* **Quantity - Quantity purchased**
* **InvoiceDate - Date of purchase**
* **UnitPrice - The target value, price of every product**
* **CustomerID - Unique Identifier for every Customer**
* **Country - Country of sales, encoded as Label**

**Skills:**

* ***Multivariate Regression***
* ***Big dataset, underfitting vs overfitting***
* ***Optimizing RMSE to generalize well on unseen data***

**Evaluation**

**What is the Metric In this competition? How is the Leaderboard Calculated ??**

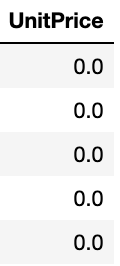
* The submission will be evaluated using the **RMSE** metric. One can use [**np.sqrt(mean\_squared\_error(actual, predicted))**](https://www.kite.com/python/answers/how-to-take-root-mean-square-error-(rmse)-in-python) to calculate the same
* This **hackathon** supports **private** and **public** leaderboards
* The **public leaderboard** is evaluated on **70%** of **Test data**
* The **private leaderboard** will be made available at the **end** of the **hackathon** which will be **evaluated** on **100% Test data**

**How to Generate a valid Submission File**

Sklearn models support the **predict()** method to generate the predicted values

You should submit a .csv/.xlsx file with exactly **122049** rows with **1** column(i.e. ***UnitPrice***). Your submission will return an **Invalid Score** *if you have extra columns or rows*.

The file should have exactly **1 column.**



***Using pandas, one can do submission\_df.to\_csv('my\_submission\_file.csv', index=False)***