About Practice Problem: Black Friday

Nothing ever becomes real till it is experienced.

*-John Keats*

While we don't know the context in which John Keats mentioned this, we are sure about its implication in data science. While you would have enjoyed and gained exposure to real world problems in this challenge, here is another opportunity to get your hand dirty with this practice problem powered by Analytics Vidhya.

This hackathon aims to provide a professional setup to showcase your skills and compete with their peers, learn new things and achieve a steep learning curve.

Data Science Resources

* **You can refer our learning path to learn more about the tools and technologies required to solve Data science problems. You can find it**[**here**](https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+Python-Final-Jan-Feb+Python-Session-1/learning_path?utm_source=practice_problem_Black-Friday&utm_medium=Datahack)**.**
* **Are you a beginner? If yes, you can check out our latest**[**'Intro to Data Science'**](https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+DS101+2018T2/about?utm_source=practice_problem_Black-Friday&utm_medium=Datahack)**course to kickstart your journey in data science.**

Rules

* One person cannot participate with more than one user accounts.
* This is proprietary dataset, you can only use for this hackathon (Analytics Vidhya Datahack Platform) not for any other reuse
* You are free to use any tool and machine you have rightful access to.
* You can use any programming language or statistical software.
* You are free to use solution checker as many times as you want.

Registration Fee

Free

Problem Statement

A retail company “ABC Private Limited” wants to understand the customer purchase behaviour (specifically, purchase amount) against various products of different categories. They have shared purchase summary of various customers for selected high volume products from last month.  
The data set also contains customer demographics (age, gender, marital status, city\_type, stay\_in\_current\_city), product details (product\_id and product category) and Total purchase\_amount from last month.

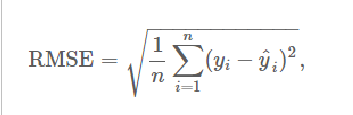
Now, they want to build a model to predict the purchase amount of customer against various products which will help them to create personalized offer for customers against different products.

Data

|  |  |
| --- | --- |
| **Variable** | **Definition** |
| **User\_ID** | User ID |
| **Product\_ID** | Product ID |
| **Gender** | Sex of User |
| **Age** | Age in bins |
| **Occupation** | Occupation (Masked) |
| **City\_Category** | Category of the City (A,B,C) |
| **Stay\_In\_Current\_City\_Years** | Number of years stay in current city |
| **Marital\_Status** | Marital Status |
| **Product\_Category\_1** | Product Category (Masked) |
| **Product\_Category\_2** | Product may belongs to other category also (Masked) |
| **Product\_Category\_3** | Product may belongs to other category also (Masked) |
| **Purchase** | Purchase Amount (Target Variable) |

Your model performance will be evaluated on the basis of your prediction of the purchase amount for the test data (test.csv), which contains similar data-points as train except for their purchase amount. Your submission needs to be in the format as shown in "SampleSubmission.csv".

We at our end, have the actual purchase amount for the test dataset, against which your predictions will be evaluated. Submissions are scored on the root mean squared error (RMSE). RMSE is very common and is a suitable general-purpose error metric. Compared to the Mean Absolute Error, RMSE punishes large errors:



Where y hat is the predicted value and y is the original value.

Please note :

* Public leaderboard is based on 30% of the test dataset, while 70% of the dataset is used for Private Leaderboard.
* The final results would be declared only on Private Leaderboard