**About Practice Problem: Big Mart Sales III**

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 access the free course on the complete EDA (Exploratory Data Analysis) and modelling for this problem using R [here.](https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+BigMS01+2018_1/about?utm_source=practice_problem_Big_Mart_Sales_III&utm_medium=Datahack)
* Are you a complete beginner? If yes, you can check out our latest ['Intro to Data Science'](https://courses.analyticsvidhya.com/courses/introduction-to-data-science-2?utm_source=practice_problem_Big_Mart_Sales_III&utm_medium=Datahack) course to get 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 Data hack 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.

**FAQs**

**1.  Are there any prizes/AV Points for this contest?**

This contest is purely for learning and practicing purpose and hence no participant is eligible for prize or AV points.

**2. Can I share my approach/code?**

Absolutely. You are encouraged to share your approach and code file with the community. There is even a facility at the leaderboard to share the link to your code/solution description.

**3. I am facing a technical issue with the platform/have a doubt regarding the problem statement. Where can I get support?**

Post your query on discussion forum at the thread for this problem, discussion threads are given at the bottom of this page. You could also join the AV slack channel by clicking on 'Join Slack Live Chat' button and ask your query at channel: practice problems.

**Registration Fee**

Free

**Problem Statement**

The data scientists at BigMart have collected 2013 sales data for 1559 products across 10 stores in different cities. Also, certain attributes of each product and store have been defined. The aim is to build a predictive model and find out the sales of each product at a particular store.

Using this model, BigMart will try to understand the properties of products and stores which play a key role in increasing sales.

Please note that the data may have missing values as some stores might not report all the data due to technical glitches. Hence, it will be required to treat them accordingly.

**Data**

We have train (8523) and test (5681) data set, train data set has both input and output variable(s). You need to predict the sales for test data set.

|  |  |
| --- | --- |
| **Variable** | **Description** |
| **Item\_Identifier** | Unique product ID |
| **Item\_Weight** | Weight of product |
| **Item\_Fat\_Content** | Whether the product is low fat or not |
| **Item\_Visibility** | The % of total display area of all products in a store allocated to the particular product |
| **Item\_Type** | The category to which the product belongs |
| **Item\_MRP** | Maximum Retail Price (list price) of the product |
| **Outlet\_Identifier** | Unique store ID |
| **Outlet\_Establishment\_Year** | The year in which store was established |
| **Outlet\_Size** | The size of the store in terms of ground area covered |
| **Outlet\_Location\_Type** | The type of city in which the store is located |
| **Outlet\_Type** | Whether the outlet is just a grocery store or some sort of supermarket |
| **Item\_Outlet\_Sales** | Sales of the product in the particular store. This is the outcome variable to be predicted. |

**Evaluation Metric:**

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

We at our end, have the actual sales for the test dataset, against which your predictions will be evaluated. We will use the Root Mean Square Error value to judge your response.

Where,  
N: total number of observations  
Predicted: the response entered by user  
Actual: actual values of sales

Also, note that the test data is further divided into Public (25%) and Private (75%) data. Your initial responses will be checked and scored on the Public data. But, the final rankings will be based on score on Private data set. Since this is a practice problem, we will keep declare winners after specific time intervals and refresh the competition.