**Project – 1: Time Series Project**

Congratulations on your new job! This time you are helping out Unicorn Investors with your data hacking skills. They are considering making an investment in a new form of transportation - JetRail. JetRail uses Jet propulsion technology to run rails and move people at a high speed! While JetRail has mastered the technology and they hold the patent for their product, the investment would only make sense, if they can get more than 1 Million monthly users with in next 18 months.

You need to help Unicorn ventures with the decision. They usually invest in B2C start-ups less than 4 years old looking for pre-series A funding. In order to help Unicorn Ventures in their decision, you need to forecast the traffic on JetRail for the next 7 months. You are provided with traffic data of JetRail since inception in the test file.

**Evaluation Metric:**

Root Mean Square Error (RMSE) is the evaluation metric for this contest

Note: Public and private split is 20:80 where first 44 days in public and next 169 days in private

*\*Note: Please refer time series folder*

**Project 2 - Big Mart Sales**

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 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

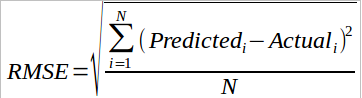
We have train (8523) and test (5681) data set, train data set has both input and output variable(s). You need to predict the sales for test data set.

|  |  |
| --- | --- |
| 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 particulat 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.

[Test File](https://datahack.analyticsvidhya.com/contest/practice-problem-big-mart-sales-iii/media/test_file/Test_u94Q5KV.csv)

[Train File](https://datahack.analyticsvidhya.com/contest/practice-problem-big-mart-sales-iii/media/train_file/Train_UWu5bXk.csv)

[Sample Submissions](https://datahack.analyticsvidhya.com/contest/practice-problem-big-mart-sales-iii/media/sample_submission/SampleSubmission_TmnO39y.csv)

Solution Checker

Top of Form

Code File

Solution File\*  
(.csv or .zip file)

Solution Description  
(max : 180 chars)\*

Add Solution

Bottom of Form

*\*Please refer folder big mart III folder for more details*

**Project 3: Retail Sales**

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 SetWe 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 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.

*\*Kindly refer retails sales folder*

**Project 4: Excel Simulation**

Welcome to the world of Assurance!

Today, you need to help Mr. Assurenaut, CEO of AssureNext. AssureNext is a radical insurance company which sits at the intersection of technology and human interaction.

AssureNext has brought some radical changes in the way they distribute Insurance. The first such call was to do away with all traditional distribution channels. So they have no bank tie ups, no agents or any online sales.So, how do they distribute Insurance to end consumers? This is where their innovation comes in. They have created a new channel called "**Robossurance"**, where they are installing smart machines in various locations.

These machines study the consumers in these locations and suggest the best Insurance plans for these customers. Currently, AssureNext is installing these machines in book stores and fine dining restaurants.

AssureNext started in 2009 and has grown multi folds in last 6 years. The company has been facing some challenges lately with manufacturing and scaling up. The customer acquisition rate has not increased as was expected in initial years. Also, the customers have grown blind to the presence of these machines over time.

Mr. Assurenaut is in a tough situation. There is a board meeting starting in 3 hours, where the board will review the future plan for AssureNext.

You need to help him create his plan and presentation.

Questions:

1. Making reasonable assumptions (and calling them out), you need to create a forecast for business (Premium received from new policies) in 2016 and 2017.
2. How many new machines would be required to be installed in order to reach 1,700 Million INR of premium from new policies in 2017
3. Assuming that you can not manufacture more than 6500 machines in a year and average premium from a policy to be INR 55,000, what should be the expected business in 2016 and 2017 (Quarter on Quarter)

Evaluation criteria

1. **Accuracy** will act as a filter. So, if your solution provides an answer in similar range as the benchmark solution from the sponsor, you would be shortlisted. Else, rejected!
2. **Methodology** - Once your solution provides similar answers, we will look at the methodology and the assumptions made in coming up with the model. Any unreasonable number would not be tolerated. **Please make sure that all your assumptions are mentioned in Assumptions sheet of the file you upload.**
3. **Quality of Automated Excel model** The objective behind creating a simulation like this is to have flexibility in business planning. Your model should take input from business users (e.g. number of machines installed & average premium per policy) and provide business outcome (in INR millions). The more flexibility you can provide to the end user to play with these numbers, the better it is.

Accuracy would be used as a filter, where as Methodology & Quality carry equal weight in deciding your score.Description of sheets in excel file is provided in "**Instruction**" sheet. Please read it before you start working.

*\*Kindly refer Excel Simulation folder*

**Project 5: Black Friday Problem**

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.

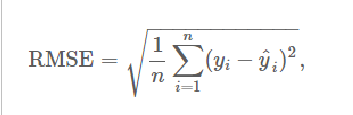
Now, they want to build a model to predict the purchase amount of customer against various products which will help them to create personalized offer for customers against different products.

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 otder category also (Masked) |
| **Product\_Category\_3** | Product may belongs to otder 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

*\*Please refer black Friday problem folder*

**Project 6: HR Case Study**

In this problem, we have provided information about various internships posted on Internshala. This includes various attributes about the internships like location, duration, start\_date of internship etc. We have also provided information about the students who have applied for the internship. These include type\_of\_institute, current\_year, academic performance of the student etc. Any student is free to apply for any internship on the portal.

While employers get high response to their posting, it is difficult to go through a high number of applications for the employers. They might need to go through high number of applications to shortlist the most relevant candidates. Hence an intelligent matching algorithm can help our users get better experience and enhance chances of*meaningful* profile matches.

Description of the files

**Internship.csv** includes the details of all the internships posted on InternShala. These details are filled by the company floating the Internship. Each row represents one internship.

**Student.csv** includes details of the students applying for the internship. These details have been filled by the student. Each row represents an experience of the student. In case the student has not filled any experience, there would be only one row containing details of student.

**test.csv** & **train.csv** include the application details (as applied by student) and the shortlist outcome

*\*Kindly refer HR Problem folder*

**Project 7: Loan Cross selling Problem**

About Company

Dream Housing Finance company deals in all home loans. They have presence across all urban, semi urban and rural areas. Customer first apply for home loan after that company validates the customer eligibility for loan.

Problem

Company wants to automate the loan eligibility process (real time) based on customer detail provided while filling online application form. These details are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and others. To automate this process, they have given a problem to identify the customers segments, those are eligible for loan amount so that they can specifically target these customers. Here they have provided a partial data set.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Data Set

Variable Description

Loan\_ID Unique Loan ID

Gender Male/ Female

Married Applicant married (Y/N)

Dependents Number of dependents

Education Applicant Education (Graduate/ Under Graduate)

Self\_Employed Self employed (Y/N)

ApplicantIncome Applicant income

CoapplicantIncome Coapplicant income

LoanAmount Loan amount in thousands

Loan\_Amount\_Term Term of loan in months

Credit\_History credit history meets guidelines

Property\_Area Urban/ Semi Urban/ Rural

Loan\_Status Loan approved (Y/N)

*\*Please refer loan cross selling problem folder*

Project 8: Loan Prediction Problem

Problem Statement

About Company

Dream Housing Finance company deals in all home loans. They have presence across all urban, semi urban and rural areas. Customer first apply for home loan after that company validates the customer eligibility for loan.

Problem

Company wants to automate the loan eligibility process (real time) based on customer detail provided while filling online application form. These details are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and others. To automate this process, they have given a problem to identify the customers segments, those are eligible for loan amount so that they can specifically target these customers. Here they have provided a partial data set.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Data

Variable Description

Loan\_ID Unique Loan ID

Gender Male/ Female

Married Applicant married (Y/N)

Dependents Number of dependents

Education Applicant Education (Graduate/ Under Graduate)

Self\_Employed Self employed (Y/N)

ApplicantIncome Applicant income

CoapplicantIncome Coapplicant income

LoanAmount Loan amount in thousands

Loan\_Amount\_Term Term of loan in months

Credit\_History credit history meets guidelines

Property\_Area Urban/ Semi Urban/ Rural

Loan\_Status Loan approved (Y/N)

Note:

1. Evaluation Metric is accuracy i.e. percentage of loan approval you correctly predict.

2. You are expected to upload the solution in the format of "sample\_submission.csv"

*\*Please refer loan prediction folder.*