**The 2017 DataHack Challenge**

**Overview**

The goal of this year’s DataHack challenge involves estimating the number of bicycle trips which will be taken by the users of a bike sharing scheme very similar to Dublin Bikes. The data has been acquired from the bike sharing scheme of a major American city. Participants are asked to build a model which can predict the total number of bicycle journeys taken on a day based on date, the weather on that date and if that date corresponded to a public holiday.

There are three data files which should be used to build your model. You are welcome to use additional data to help augment this data. The training data files are

* **Bikes.csv**

This data file contains two columns

1. Date: The date in question
2. The number of bicycle journeys taken on the corresponding date

* **Weather.csv**

This data file contains multiple columns pertaining to the weather

1. Date: The date in question
2. PRCP: This is a measure of rainfall/precipitation
3. SNOW: This is a measure of snowfall
4. TMAX: This is a measure of the maximum recorded temperature.
5. TMIN: This is a measure of the minimum recorded temperature.
6. TAVG: This is the average recorded temperature.
7. AWBD: This is a measure of average wind speed

* **Holidays.csv**

This data file contains a list of public holidays in the state of the bike sharing company with two columns.

1. Date: The date in question
2. Holiday: the name of the public holidays

Please note that not all of these columns may be important to your final model.

**Suggested Approach**

When approaching a machine learning problem such as the one described above there is a series of steps which are best followed

1. Open the data in your editor of choice and browse through it to understand it.
2. Load the data into R, Python or your data analysis program of choice and review the integrity and quality of the data. Check for missing data or corrupt data.
3. Review the statistics for each column and identify potential outliers which may need to be removed.
4. Create a single table containing all the data organised by date.

You are now ready to begin building your machine learning model. While there are other data pre-processing steps which you may improve your model the steps described above will allow you to quickly start iterating on your model.

As this machine learning problem involves estimating the number of bicycle trips taken per day it is a “Regression” type problem. There are multiple different approaches which you can take to begin building your model. If you are unsure of how to proceed please ask one of the Data Scientists for help!

**Scoring**

You model will be scored throughout the day using an automatic scoring algorithm. In order to have your model scored you must predict the number of bicycle journeys which will be taken based on the data in the Testing.csv dataset. This dataset contains all the relevant columns which you will need to score your model. In order to submit your model for evaluation simply do the following

1. Create a “Results.csv” file. This file should contain two columns
2. ID (the ID given in the Testing.csv file)
3. Total (Your estimated number of journeys)

2. Add your “Results.csv” file to your Dropbox folder

The automatic scoring algorithm will not work if you name the file or label the columns incorrectly. It will also not work if your model doesn’t produce an estimated number of journeys for each ID. If you have an issue please double check these things before letting a member of the DataHack team know about your difficulty. The scoring algorithm will run every minute so it may take a while for your model to be scored. Once scored, a file with called “Score.txt” will appear in your Dropbox folder. The score is calculated as follows

An accurate model will have a lower score than an inaccurate model. Participants will be able to see how they are doing on the Scoreboard.

At the end of the day participants will be asked to score the “Final\_Test.csv” file in the same way. This Final\_Test result will be used to identify the most accurate model in order to avoid the potential of the winning model having simply overfit to the Testing.csv data. Due to this the final standings may be different to what was visible on the scoreboard before the end of the competition. However, the scoreboard will still give a good indication of overall model accuracy.

**Best of luck and if you have any other questions please ask of the AIB Data Scientists!**