2021 Chevron Rice Datathon:

Predicting Hot Dog Sales

## Description

Every year thousands of hot dogs are toss into the garbage by Fuel Station Managers due to sanitary restrictions on how much time they are kept warm. Having a robust cook plan for every day would grant the stores a sales forecast and a money saving tool. During this challenge, the participants are requested to create a cook plan for certain buckets of time.

Stores are open from 8 am to 8 pm every day, including weekends.

The 3-hour time buckets provided are:

1 = 8am to 10 am

2 = 11am to 1 pm

3 = 2pm to 4 pm

4 = 5pm to 8 pm

When planning the logistics of the store, the store managers need to know with time in advance how many hot dogs are they going to sell during that month. In this Data Science Challenge, you are tasked with developing a model to predict:

*The Gross Sold quantity of hot dogs for different days and buckets for some stores.* ***Please keep in mind that your model must predict an integer number of hotdogs to be sold, there is no such thing as 7.5 hot dogs.***

## Data

**DISCLAIMER**: The data provided for this challenge can be used only to develop a model for the challenge and should be discarded after the end of the challenge.

The challenge involves several files for your project. The files required to build and test your predictive model will be made available at the start of the Datathon. However, 30 minutes before the judging period starts, we will also release a scoring file that contains additional data for which you will run your final model and provide corresponding predictions.

The following files will be available on Piazza at the start of the Datathon :

* **training.csv**: contains raw data used to build, train and test your model. The dataset is listed by StoreNum, day of the year (1 through 365), 3 Hour Bucket (1 through 4) and the Gross Sold Quantity (an integer your model will predict). It also contains information regarding the store selling the hot dogs, like location and if it has a car wash or not.
* **data\_dictionary.xlsx**: provides description of the variables in training.csv
* **scoring\_format.csv**: shows the format of the scoring file (scoring.csv, released 30 minutes before the Datathon ends). This is a sample of rows from training.csv, but the real scoring.csv file will comprise of completely new rows of data.
* **submission\_format.csv**: gives the required format of the model predictions file obtained by applying your model to the scoring.csv file. Note: the columns *StoreNum, dayOfTheYear,* and *3HourBucket* in this file corresponds to the *StoreNum, dayOfTheYear,* and *3HourBucket* in the scoring.csv file below and ***the rows must be submitted in this same order***!

The following file will be available on Piazza 30 minutes before the Datathon judging begins:

* **scoring.csv**: used to generate model predictions required for us to score your model accuracy
* **submission\_file.xlsx**: used to submit model predictions and calculate model accuracy. Input your model predictions into the “Predictions” sheet. After predictions are inserted, your model’s predictive accuracy will calculate automatically so ***do not make any changes to the “RMSE” sheet***.

## Project Submission

Please submit your Datathon project to devpost for judging. At minimum, the submission should include a link to a github repo that contains:

* your project code
* a summary of your methodology and findings
* submission\_file.xlsx: the file in which you inserted the predictions outputted from running your model on the scoring.csv dataset

Due to the specificity of the submission\_file.csv, do not make any changes to the order of the rows on the “Predictions” tab. Make sure predictions are placed with the correct corresponding observation.

## Evaluation Criteria

Projects will be evaluated holistically based on model accuracy the criteria outlined by Rice Datathon:

* **Technical Difficulty**: We are looking for technically advanced solutions to difficult problems that make use of a diverse set of modeling and data science techniques. That being said, if you can solve a challenging problem with a simple solution, we will be very impressed!
* **Analysis & Exploration**: We are looking for projects that take time to analyze and explore the nuances of whatever data they are working with.
* **Creativity**: We are looking for original ideas or new angles on existing ideas.”
* **Predictive Accuracy**: predictive accuracy will be measured by the Root Mean Squared Error (RMSE) in Gross Sold Quantity predictions for the stores in the scoring.csv file:

## where is the actual Gross Sold Quantity corresponding to the -th row in scoring.csv, and is the corresponding model prediction.

Other areas that can be looked at include, but are not limited to feature generation, feature selection, and model selection/building.

## Rules

1. Contestants must respect the privacy of the data and remove it from their computers upon completion of the competition.
2. Contestants' solution must be a model that is repeatable, adjusting model results manually (including “arbitrary” factors/constants) to tune model predictions is not permitted.
3. Contestants' prediction models can use only that data from the training.csv dataset provided to build the model.