Data Science Work Sample

Root Insurance Co.

This task consists of building a predictive model, including implementing feature extraction on real data, and is intended to help us assess your capabilities as a data scientist. Your submission will be evaluated based on the following (which should be clear from the notebook or code/report)

- **Exploratory data analysis**: understanding the data set, including quality and relevance of different predictors, etc.
- **Programming ability and code quality**: design decisions (such as for feature extraction), algorithmic efficiency, and readability of code
- **Model building**: choice of appropriate models, cost functions, hyperparameters, etc.
- **Model diagnostics**: evaluation of model performance, in- and out-of-sample performance, and how this informs model building iteration process
- **Technical communication**: clarity and conciseness of submitted report or notebook markdown, and conveying the relevant findings and metrics effectively

This task is not intended to take an inordinate amount of time, nor is it expected that the results be perfect. We simply wish for the candidate to make their best effort at coming up with a reasonable solution to the problem. A typical suggested time to spend on the work sample is around 4 hours.

Please do not share the data, instructions, or your implementation with any other party.

1 Task

For this work sample, a few things are provided:

- A training dataset consisting of:
  - A dataset of N_{train} "trip" csv data files: timeseries data recorded from car trips including the vehicle speed and heading at different points in time
  - A data matrix $\boldsymbol{X} \in \mathbb{R}^{N_{train} \times K}$ of features about each trip
  - A target vector $\boldsymbol{y} \in \mathbb{R}^{N_{train} \times 1}$ indicating whether each trip is "interesting" or not
- A test dataset (similar to the training dataset, but with N_{test} records) consisting of just the trip data and the features, but no labels

The objective of the work sample is to augment $\boldsymbol{X}$ by adding up to two new features $\boldsymbol{x_{a1}}, \boldsymbol{x_{a2}}$, and use the final feature set $\boldsymbol{X_a} = [\boldsymbol{X}, \boldsymbol{x_{a1}}, \boldsymbol{x_{a2}}] \in \mathbb{R}^{N_{train} \times (K+2)}$ to fit a model $\hat{\boldsymbol{y}} = f(\boldsymbol{X_a})$; this model will then be used to to predict $\boldsymbol{y}$ on the unlabeled test data.

2 Data

The prepared files include:

- `trip_data_train.zip`: a zipped folder containing the relevant csv trip data, where each trip has a series of records containing:
  - `time_seconds`: the time in seconds since the start of the trip (float)
  - `speed_meters_per_second`: the speed in m/s of the vehicle (float $\in [0, \infty)$, invalid values will be indicated with NaN)
  - `heading_degrees`: the angle in deg of the vehicle's travel relative to north (clockwise positive, float $\in [0, 360)$, invalid values will be indicated with NaN)
- `model_data_train.csv`: a csv file containing:
  - a column `filename` to join to the file names in `trip_data_train.zip`
  - the data matrix $\boldsymbol{X}$ as columns `feature1`, ..., `featureN`
  - the vector $\boldsymbol{y}$ as column `y`
- `trip_data_test.zip` and `model_data_test.csv`: similar trip data but without the label column `y`.

3 Instructions

You should implement the code to perform exploratory analysis, data transformations, model fitting, and evaluation in a script or notebook. Using Python 3 or R is highly preferred, and results delivered in a Jupyter or R-markdown style notebook will make evaluation of your submission much easier. Please try to prune the content of your submission to the "relevant findings" and consolidate analysis, model selection, etc. to a handful of representative figures. If a script rather than a notebook is used, the result can be delivered with an accompanying PDF or other file to include figures as needed.
The additional features to augment the data matrix should be the following (**note the candidate may chose to implement one or both of the features, as time permits**):

- x_{a1}: the count of stops in the trip [int]
- x_{a2}: the count of turns (greater than 60 degrees) [int]

It should be noted that these are not exact definitions, and the candidate is expected to be judicious in deciding for themselves what sort of logic appears to be "reasonable". For turns and stops, consider events within 3 seconds of one another to be the same event, and exclude events lasting shorter than 3 seconds. A left turn followed immediately by a right turn (or vise versa) should be treated as two separate turns. To help with building the feature extraction, the trip data corresponding to `filename = "0001.csv"` should have approximately:

- 1 stop
- 9 turns

Because some of these features don't have exact definitions, perfect agreement is not expected, but better agreement with realistic physical behaviors will result in better correlation with the target.

The (discrete binary) target variable $y_i \in \{0, 1\}$ represents whether the i^{th} trip is "interesting" or not, and the objective of the work sample is to fit a model (using the training data) that can make an accurate prediction $\hat{\boldsymbol{y}} = f(\boldsymbol{X_a})$ of $\boldsymbol{y}$. The model will then be used to generate predictions on the provided test data. The candidate should deliver the test predictions in csv form like:

```
filename ,prediction
0000.csv ,0
0001.csv ,1
...
```

where `filename` is relevant filename from `trip_data_test.zip` and `prediction` is a binary integer, 0 or 1.

Further, a small write-up should be included that assesses the performance of the feature extraction (in terms of computational complexity) and the model (in terms of evaluation metrics). An estimate of the out-of-sample accuracy, recall, and precision should be provided in the writeup. Please do not feel that the write-up needs to be extremely professional or beautifully formatted; rather, it just needs to be clear enough to convey the approach taken and highlight the performance metrics. If a notebook is delivered, the write-up can be included as a section of the notebook.

You may also discuss assumptions you made while working on the problem, steps you took to overcome certain pitfalls, things to try if you want to improve current results, or anything else you think is relevant, but please keep your report and discussion relatively brief and focused on the key findings.

4 Output

The deliverable should consist of two or three parts:

- The notebook or script containing the code to solve the problem
- The csv file with the test data predictions, formatted as specified in Section 3.
- *optional*: supplementary PDF report to accompany non-notebook submission, containing description of EDA, feature generation, modeling, evaluation, and any figures

Once finished, please create a zipped file including any of your work, or anything our reviewers will need. Please rename the document in the format "(first initial)(last initial)_worksample". For example, the Candidate Experience Manager's name is Bryan Croft, so his work sample would be "bc_worksample".

Outside of your initials on the zipped file, please do not include your name or any personally identifiable information in your submission. We anonymize work samples when we review them to reduce the chance for bias in the review process.