Homework #2

due Wednesday, Sep 25 (end of class)

(work in pairs – groups available in Canvas)

In this assignment we are revisiting the "bobblehead" problem, and analyze it using the latest techniques learned in the class and by using Python. You will need to submit a Jupyter notebook with the answers to the questions in this assignment via Canvas. Make sure that the name of both students working on this assignment is listed in the report and Python notebook. *Name your notebook file using all the last names of the students in the group separated by "_" (underscore) – example "lastname1_lastname2.ipynb".*

Download the following two files from Canvas:

- dodgers_training.csv
- dodgers_testing.csv

You are asked to build regularized linear models to predict the game **attendance** based on the **month, day_of_week, temp, skies**, and **bobblehead** input attributes.
You need to perform the following data preprocessing operations:

- Encode the categorical variables using the one-hot encoding.
- Make sure that in the process both the training and testing datasets have the same data columns (input attributes).
- Standardize the new features by removing the mean and scaling to unit variance.
- Using the training dataset, train 100 L2-regularized linear models corresponding to 100 regularization coefficients evenly spaced between 0.1 and 1000. Use the leave-one-out cross-validation.
- Similarly, perform the same operation but now using L1-regularization.
- Train also a linear model without regularization.

Respond to the following questions:

1. [10pts] List the means and standard deviations of the encoded attributes before scaling.

2. [15pts] What is the best L2 regularization coefficient? Provide the corresponding linear coefficients.
3. [15pts] What is the best L1 regularization coefficient? Provide the corresponding linear coefficients.
4. [15pts] What are the predictive attributes selected as a result of L1 regularization?
5. [15pts] For the models corresponding to the best L2 and L1 regularization coefficients list the following root-mean-square-error (RMSE): (1) RMSE on the training set, (2) expected prediction RMSE obtained during the cross-validation, (3) RMSE on the testing set. For the linear model without regularization list the following RMSE: (1) RMSE on the training set, and (2) RMSE on the testing set.
6. [15pts] What observations can you make based on the above RMSE?
7. [15pts] Predict the attendance on a clear Monday in June when the expected temperature is 72 for all three models with and without bobbleheads. Does bobblehead promotion have an impact on the attendance?

BONUS [10pts] – if you accomplish the same thing in Orange and submit your workflow.