Project 3 Presentation

Requirements:

-Prepare a **15-minute** data deep-dive or infrastructure walkthrough that shows machine learning in the context of what we've already learned.

-(OPTIONAL) Host application using Heroku or a tool of your choice

Presentation Outline:

- We employed a Kaggle dataset that provided ratings of ice cream flavors from 4 of the top brands.
    - Breyer's
    - Ben & Jerry's
    - Haagen-Dazs
    - Talenti
- The data was comprised of 21,674 reviews of 242 ice cream flavors rated on a 5-star Likert scale along with comments entered by each person submitting a review (reviews.csv). A summary of the ratings for each ice cream flavor and their ingredients was provided in a separate data file (products.csv).
- We initially gave some consideration of doing sentiment analysis on the comments in the reviews.csv datafile but were concerned we would not accomplish this in the time frame allotted.
- We decided to employ the summary dataset with the intent of creating a model to predict the ratings based on the component ingredients.
- The data was easy to load and view in Excel.
- First Challenge:
    - The ingredients list for each ice cream flavor was a single string with as many as 72 individual ingredients in the string separated by commas.
    - We used both Excel and One-Hot encoding to separate the ingredients into individual sequential columns.
    - There are a total of 414 unique ingredients.
- We wanted to create a website where a user could submit a list of ingredients they chose, and our model would predict an average rating.
- Second Challenge:
    - Presenting the user with a list of 414 ingredients to choose from would not be practical.
    - We sorted the list of ingredients by frequency of use and tried to find a cutoff point, i.e. eliminate ingredients with frequency of use below 10.
        - The problem with this approach was that several choice ingredients had low frequency of use values. We felt this might hamper our model's ability to predict a rating.

    - We wound up hand-selecting the ingredients to include in our final ingredients list by group vote as we went through the unique ingredients list one by one.
    - This list (final_ingredients.csv) had 64 unique ingredients in 4 categories: Basics, Dairy, Flavors, and Toppings.
- We recalculated the frequency of use of the each of the 64 ingredients and found minor differences, but the rank order of ingredients remained the same as in the full dataset.
- We created a file that included only the ratings and the 64 selected ingredients to use in our machine learning portion.
- We created a second file that included only the ratings and all 414 unique ingredients to use in our machine learning portion as comparison.
- One important thing to note here is that our 'label" is a linear variable while our "features" are all categorical values.
    - The dataset we had did not share the actual recipe for each ice cream flavor, it only listed which ingredients were present.
- Our first attempt was to use the linear regression kernel to generate a model.
    - Using the data for the 64 selected ingredients, this model did not perform very well with an R^2 of 0.507 for the training set and a terrible R^2 for the test data.
    - Using the data for the full 414 ingredients, the model performed even worse, surprisingly.
- We then used the random forest kernel running the RandomForestRegressor.
    - Using the data for the 64 selected ingredients, this model performed respectably with the training data ($R^2 = 0.83$) but not as well with the test data ($R^2 = -0.05$). This may show a dependence on larger input arrays?
    - Using the data for the full 414 ingredients, the model again performed worse.
- The Random Forest kernel includes an attribute (df.feature_Importances) that allows us to graph the relative importance or weight of each feature in our model's predictions. Out of curiosity when ran this on our model. It shows some general principles of how features work in a model.