CS148 Project 3

Due Monday March 14, 2022 before Midnight PST via Gradescope

Introduction:

Congratulations CS148! You've been officially hired as consultants to work with Munchies* (not a real company!), one of the largest and fastest growing Cannabis Brands in the world. Munchies is eager to develop their datascience capabilities and you have been recruited to help them develop some of their initial predictive models.

Challenge:

You will serve as consultants to Munchies. Having been provided with a complete set of cannabis sales data going back to 2018, along with detailed product descriptions for all brands sold in California, you will be asked to:

1. Develop a predictive model to help forecast product sales
2. Conduct an analysis to determine the key factors that likely impact the success of a product and based on the data analysis propose potential growth areas to the company

This project will include both a structured component, where much like Projects 1 and 2, you will be given a specific set of instructions to complete. There will also be an unstructured component where we will ask you to experiment with your own approaches to see if you can maximize your own results.

One important thing to note: **We will not tell you precisely what to model.** The data provided includes a variety of possible target labels for you to choose from. Part of your work is to figure out precisely what you'd like to try and model.

Project Overview:

This project will consist of two discrete components. Full credit for Project 3 will require your completion of all components. Specifically you will be asked to produce or submit to the following:

- **Report:** A report documenting your work on the project and your findings
- **Coding Project:** Follow the steps detailed below on a Jupyter Notebook and output your code fragments along with the results

Final Deliverables:

- PDF with the output of Jupyter Notebook (submitted via Gradescope) and the Final Report (Submitted via Gradescope)

Timeline:

- Project will be released on Feb 28th
- Project will be due on the monday of Finals week (Monday, Week 11) on **March 14, before Midnight PST**

Project Requirements:

Specific Coding Requirements:

1. **Merge Datasets and Effectively Link information -** Useful information for this project will come from disparate datasets. You will need to effectively merge them into a single dataframe for analysis
2. **Develop basic Time Series Feature Extraction Plan -** develop a series of standard timeseries features to augment your dataset and enable timeseries predictive models.
3. **Create additional data feature engineering plan and implement it (no need to pipeline this)** Determine and execute a plan to fully pre-process your data for modeling and then execute it. Specifically:
    1. Determine which fields to retain and which to drop.
    2. For those you retain, determine a categorization strategy
    3. Determine an imputation strategy (you should choose more than one imputation method depending on the specifics of your data
    4. Augment at least one feature, ideally a feature cross, or non-linear transition e. Determine a strategy for scaling features
4. **Implement a basic Linear Regression predictive model for statistical hypothesis testing** - With your newly pipelined data find and interpret important features (e.g. using regression and associated p-values). If there are any collinearities be careful when incorporating them into the regression.
5. **(Extra Credit): Implement Principle Component Analysis (PCA)** - Since your resulting dataframe is likely to be high-dimensionality, employ PCA to reduce the complexity of your dataframe. Test the effect of different levels of principle components on your linear model and see if you can optimize performance. Note: since we haven't covered this in class. This action item will be extra credit.
6. **Employ an ensemble method to your predictive model exercise -** Leverage an ensemble learning method to generate an optimized prediction model
7. **Cross-Validate your training results -** Employ K-Fold Cross-validation to your training regimen for both ensemble and single regression models. (Optional: employ a stratifiedshufflesplit as well to ensure equitable distribution along a key parameter)

8. **Employ a GridSearch method to optimize your parameters -** Leverage gridsearch or an equivalent parameter tuning approach to optimize parameters to your predictive model (Note: you can likely merge the gridsearch and cross-validation steps into one single run!)
9. **Experiment with your own custom models and report out your highest performing model -** For this part of the project you have free range to employ any of the tools you've learned in class, along with any additional tools or techniques you research independently to see how you can do.

Report Requirements:

You will be asked to submit a brief report accompanying their project. There is no specific length or formatting requirement. Points will be deducted for incomplete or unprofessional reports.

The report will be expected to contain the following sections:

1. **Background/Introduction:** Use the accompanying information provided by Munchies, along with your own industry research, to better explain the domain challenges. Also please specify precisely what you are trying to model/predict and why you made that decision.
2. **Methodology:** Incorporate requirements 2, 3, 8 and, 9 from the coding requirements into a general description of the work that you have done on this project. Provide an explanation or justification for why you chose the data and the methods you did, and also detail any experiments you ran and the results
3. **Results:** Report out your results from coding requirements 4,5,6,7,8, 9 on the performance of your predictive model. Additionally, report out your findings on key indicators for the likely success of a new product launch in the current market based on the statistical analysis performed in step 4
4. **Discussion:** Provide context to the results you've obtained. Additionally, provide a set of recommendations to Munchies for how to leverage your findings along with next steps for analytic work