CS148 Final Project

Due Thursday Dec 2nd, 2021 at Midnight PST via Gradescope

Introduction:

Congratulations CS148! You've been officially hired as consultants to work with Cookies, one of the largest and fastest growing Cannabis Brands in the world. Cookies is eager to develop their datascience capabilities and you have been recruited to help them develop some of their initial predictive models.

Background:

Cookies was founded in 2012 by Gilbert Anthony Milam Jr., better known by his stage name Berner, and his partner Jigga, Bay Area cultivator and breeder. The cofounder leveraged a unique confluence of emergent genetic engineering, internet culture, and strong associations with the music industry, to be one of the first companies to establish an identity and streetwear company that has come to dominate the industry.

Today, Cookies is one of the most respected and top-selling cannabis brands in California and is globally recognized, amassing a stable of over 50 cannabis varieties and product lines including indoor, outdoor and sungrown flower, pre-rolls, gel caps and vape carts. With two flagship Cookies stores in Los Angeles on Melrose and Maywood and a third location in Redding, Cookies' overall vertical integration and seed to sale business allows for complete quality control at every step from cultivation and production to retail experience.

In 2015, the brand's hip-hop credibility effortlessly expanded Cookies into streetwear and today offers a range of products for both men and women in the apparel and accessories categories as well as a curated selection of smoking supplies.

Challenge:

You and your partner (yes you will be allowed to work in groups of up to 2!), will serve as consultants to Cookies. 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.

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

Timeline:

- Project will be released on Nov 5th
- Project will be due before the last class of the quarter (Thursday, Week 10) on **December 2, before Midnight PST**

Collaboration Policy:

- Project work can be completed individually or as a group effort
- Groups of **up to two** members will be allowed
    - ○ There will be no grading scale for group work (i.e., no difficulty adjustment for groups) so group work is encouraged
- **Groups should submit only once** to Gradescope, but make sure to tag both participants to ensure credit for their work.

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. **Run some basic statistics on your variables including correlations with labels and report findings** - Particularly once you employ PCA and other 'black box' methods, the descriptive power of any of your features will effectively disappear. Still you want to report out meaningful correlations toCookies to help them flag key indicators they can employ (this step will also be helpful for you in flagging potential co-linearities).
4. **Create additional data feature extraction plan and implement a comprehensive pipeline to execute it -** Determine and execute a plan to process your data for modeling and then implement a pipeline to 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
5. **Implement a single pipeline to execute this transformation**
6. **Document your data strategy in your report.** Provide an explanation or justification for why you chose the data you did, and also detail any experiments you ran and the results
7. **Implement a basic Linear Regression predictive model** - 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.
8. **Implement Principle Component Analysis (PCA)** - Since your resulting dataframe is likely to be high-dimensionality, employ PCA to reduce the complexity of your dataframe
9. **Employ an ensemble method to your predictive model exercise -** Leverage an ensemble learning method to generate an optimized prediction model
10. **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)
11. **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)
12. **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:

Each team will be expected to submit a report accompanying their project. There is no specific length or formatting requirement, however this report will be shared with ..., and therefore is expected to be professionally produced. Points will be deducted for incomplete or unprofessional reports.

The report will be expected to contain the following sections:

1. **Executive Summary:** Single-page highlevel summation of the work done and key findings
2. **Background/Introduction:** Use the accompanying information provided by Cookies, along

with your own industry research, to better explain the domain challenges

3. **Methodology:** Incorporate requirements 2, 6, 9 and 10 from the coding requirements into a general description of the work that you have done on this project
4. **Results:** Report out your results from coding requirements 3, 7, 9, 11, 12 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
5. **Discussion:** Provide context to the results you've obtained. Additionally, provide a set of recommendations to Cookies for how to leverage your findings along with next steps for analytic work
6. **Conclusion:** concisely summarize the work done on the project