Pace – Dataiku Weekend Hackathon 2022

Business Problem

In a world where people are divided in their beliefs and opinions on almost everything, the one thing they will agree on is a love for chocolate. When it comes to chocolate, it is not about whether but how much people like it. In this hackathon, you will examine the characteristics of chocolate that determine how likable it is. Thus, there are two goals of this hackathon

- Explanation: Describe the characteristics of chocolate that make it likable (or delicious)
- Prediction: Accurately predict the rating of chocolate based on its characteristics

Here are some points to consider as you address the problem

1. Models: There are numerous predictive models ranging from linear regression to gradient boosting methods. While predictive models differ in many ways, a key issue relevant to the problem is the tradeoff between model accuracy and interpretability. Models that generate good predictions tend to perform poorly at explaining the relationships between the predictors and outcome and vice versa. The choice of your model should reflect this tradeoff.

2. Predictors: A model can only perform as well as the predictors going into it. You are given a set of over thirty variables that describe each chocolate along a set of characteristics such as cocoa percent, creaminess, ingredients. It is important to give some thought to which predictors go into the model. Leaving out a variable could compromise the predictive power of the model. On the other hand, including poor predictors into the model may end up hurting performance of the model out of sample.

3. Features: Generally speaking, transforming raw variables into features helps boost predictive power of a model. It is best if such transformations are made based on understanding of the data. Variable transformations done solely for the purpose of improving predictive accuracy may make it hard to explain the relationships between the predictors and outcome.

4. Generalizability: The predictive power of a model should be judged on its performance on unseen data. For this reason, we recommend following the convention of splitting the data into a train and test sample retaining about 75% of the data in the train sample. The model should be estimated on the train sample and the best model be evaluated against the test sample.

5. Outcome metric: There are a number of metrics for predicting rating of a model including those popular in statistics (e.g., R2, AIC) and those derived from prediction error (e.g., RMSE, MSE). We recommend using the Root Mean Squared Error (RMSE) as we will be judging predictions based on RMSE.

Data

The dataset describes over 2000 chocolates based on the source of the coffee bean, company, percent cocoa, ingredients, most memorable characteristics, and rating. We are interested in learning about the characteristics of chocolate that make it likable (or delicious) and constructing models to predict rating.

Variables

- id: Unique identifier
- company_location: location of company
- country_of_bean_origin: country of origin of coffee bean
- cocoa_percent: Cocoa percent (% chocolate)
- rating: Rating of chocolate
- 20 Memorable Characteristics Variables: Yes or No
    - sweet
    - nutty
    - cocoa
    - roasty
    - earthy
    - creamy
    - sandy
    - fatty
    - floral
    - intense
    - spicy
    - sour
    - vanilla
    - fruit
    - molasses
    - woody
    - sticky
    - coffee
    - rich
    - dried.fruit
- 7 Ingredients: Yes or No
    - ingredient_Beans
    - ingredient_Sugar
    - ingredient_Sweetener
    - ingredient_Cocoa_Butter
    - ingredient_vanilla
    - ingredient_lecithin
    - ingredient_salt
- number_of_ingredients: Number of ingredients in the chocolate

Files

You are provided with two data files, (a) data.csv and (b) scoringData.csv. The first file, data.csv, is meant for training your model while the latter, scoringData.csv is for applying your training model to generate a set of predictions. For this reason, scoringData.csv does not contain the variable, rating. We will compare your predictions of rating to the true value of rating to judge the quality of your predictions.

Analysis Software

You may use any analysis software you see fit to address the above problem. Here are a few popular ones

- Data Cleaning or Preparation: Excel, SQL
- Analysis: SAS, SPSS, R, Python, RapidMiner
- Communicating results: Excel, Powerpoint, Tableau, PowerBI, R, Python

Our partner, Dataiku has generously offered use of their platform for the Hackathon. Dataiku is an integrative platform that contains a number of tools for cleaning, analyzing and visualizing in a single ecosystem. For those averse to coding, Dataiku's platform makes it possible to do all the things needed for this hackathon with minimal or no code. You can access the Dataiku instance here. Your username is noted in the first column of the teams spreadsheet and your password is Ht6761w03GfIn0y9WC. Be sure to change your password after you login and share the new password with all your team members. Here is the Zoom Recording (passcode is ^6v0HtCr) of the session done by Chris and you can learn more from free online courses on Dataiku Academy.

Location

You may participate in the Hackathon from anywhere you want to. It is up to you to figure out how to connect with and work with your team members.

The Hackathon is limited to Pace University students.

Submission Guidelines

1. Predictions

Submit predictions for your scoring data as a CSV file with only two columns, “id” and “predicted_rating”. The submission file should have as many rows as the scoring data. See example_submission.csv for a sample of what the file should look like.

2. Report

The report should describe the research problem(s), exploratory data analysis, data tidying, methods used, analysis, findings, and recommendations. The body of the report should be at most 1000 words (i.e., about two pages). Charts and tables are encouraged and are not a part of the word count. Supporting code and data (original and potential enrichments) should be uploaded to a cloud resource and linked to in the report.

Report Rubric

1. Conceptualization of research problem: Clearly state the questions the analysis sought to address. What is the motivation/importance of these questions? What could be the possible impact of their answers?
2. Extent of data exploration: Use a variety of graphical representations and summary statistics to gain meaningful insights from the data.
3. Quality of data tidying and feature engineering: Take appropriate measures to clean and transform the data, as well as generate new features via the existing data and/or enrich the data with external data sources.
4. Breadth of analytical techniques examined: Use multiple analytical models. Tune models to optimize out of sample error. Justify choice of models trained.
5. Criteria used for assessing model(s): Choose an appropriate metric for evaluating the model.
6. Interpretation of results: Concise summary of the primary takeaways as they relate to the research questions of interest. Support your findings with graphical/numerical results, findings from the model(s) and relevant variable and/or parameter estimate interpretations.
7. Conclusions and recommendations address research problem: Summarize the primary takeaways as they relate to the research questions of interest and provide advice for clients. Include a discussion on challenges, questions for further research and limitations of your completed analysis.

3. Slide Presentation

Just as all data scientists should be great technicians, so too should they be great communicators. For this challenge, you are expected to deliver final results and recommendations via a 5-minute presentation. Record your presentation via Zoom (or similar recording software) and share a link to the recording.

The team presentation should convey the work described in the report but in a succinct and persuasive manner. It should focus on the data analysis steps that led to final results, and not on the wrong turns and dead ends. Students may use any visual aid including slides, presentable code (e.g., Jupyter Notebook, Knit R Code), and data visualization software (e.g., Tableau). Wherever relevant, convey content using charts or pictures.

Structure the presentation to include the following elements:

1. Research Problem: Describe the questions the analysis sought to address.
2. Exploration and Tidying: Highlight insights from data exploration. Discuss data cleaning and feature engineering conducted.
3. Analysis: Discuss the analytical technique used. If more than one analytical technique was examined, discuss the techniques and offer a justification for the predictive model ultimately used.
4. Findings: Summary of the primary takeaways as they relate to the research questions of interest. This could include a numerical and/or graphical summary of key findings from the selected model(s); relevant variable and/or parameter estimate interpretations.
5. Recommendations and Conclusion: Summary of the recommendations in a concise manner