Case Study Project - Coffee Shops

We have included a sample solution at the end of this document. Check it out to see what a passing solution to this case study looks like.

The dataset for the exam is only available in a DataCamp Workspace. However, the dataset for this sample can be downloaded from here.

Company Background

Java June is a coffee franchise looking to expand its business to a new market. Their strategy of rapid and sustainable growth is to get the most number of reviews in one year after a new coffee shop opens. Based on their data, all coffee franchises in other markets can get over 450 reviews on average after one year.

Customer Questions

The marketing manager has asked you to answer the following:

- Can you predict whether a newly opened coffee shop can get over 450 reviews based on its characteristics?

Dataset

The dataset contains information about coffee shops after 1 year of opening in this new market. The data is available in a DataCamp Workspace, which you can find from the certification dashboard.

The dataset needs to be validated based on the description below:

Column Name	Criteria
Region	Character, one of 10 possible regions (A to J) where coffee shop is located.
Place name	Character, name of the shop.
Place type	Character, the type of coffee shop, one of “Coffee shop”, “Cafe”, “Espresso bar”, and “Others”.
Rating	Numeric, coffee shop rating (on a 5 point scale). Remove the rows if the rating is missing.
Enough Reviews	Binary, whether the number of reviews is over 450 or not, either True or False.
Price	Character, price category, one of 3 categories.

Delivery option	Binary, describing whether there is a delivery option, either True or False.
Dine in option	Binary, describing whether there is a dine-in option, either True or False. Replace missing values with False.
Takeout option	Binary, describing whether there is a takeout option, either True or False. Replace missing values with False.

Submission Requirements

1. You are going to create a written report summarizing your findings. Use the project task list provided below for guidance in the tasks you should complete and information to include in the report.

2. You will need to use DataCamp Workspace to complete your analysis, write up your findings and share visualizations.

3. You must use the data we provide for the analysis.

4. Use the grading rubric provided below to check your work before submitting the report.

Project Task List

Data Validation

1. Check the data against the criteria in the data dictionary.
2. Describe what you found in the validation process. Have you made any changes to the data to enable further analysis?

Exploratory Analysis

1. Explore the characteristics of the numerical and categorical variables.
2. Create at least two data visualizations to demonstrate the characteristics of variables.
3. Create at least one data visualization to demonstrate the relationship between variables.
4. Describe what you found in the exploratory analysis. Have you made any changes to those variables to enable model fitting?

Model Fitting

1. Describe what type of machine learning models are suitable to address the problem.
2. Choose and fit a baseline model.

3. Choose and fit a comparison model.
4. Explain the reason for choosing the two models above.

Model Evaluation

1. Evaluate the performance of two models by appropriate metrics.
2. Compare the evaluation results between two models and describe what that means for addressing the business problem.

Grading Rubric

You will be graded against the following criteria. You must pass all criteria to pass this part of the certification.

Domain	***Description***	**Sufficient**	**Insufficient**
Data Validation	Assess data quality and perform validation tasks	Has validated all variables against provided criteria and where necessary has performed cleaning tasks to result in analysis-ready data.	Has not conducted all the required checks and/or has not cleaned the data. May have removed data rather than performed cleaning tasks.
Data Visualization	Create data visualizations in coding language to demonstrate the characteristics of data and represent relationships between features.	Has created at least two different types of data visualization that highlight characteristics of variables after validation. Has created at least one visualization that shows the relationship between two variables. Has used visualizations that support the findings being presented.	Has used the same visualization throughout. Has not included graphics to represent single variables and relationships. Has not used visualizations that support the findings being presented.
Model Fitting	Implement standard modeling approaches for supervised or unsupervised learning problems	Correctly identified the type of problem (regression, classification or clustering) Has selected and fitted a model for that problem to be used as a baseline. Has selected and fitted a	Has incorrectly identified the type of problem. Has not fitted a baseline model or has used a model for the wrong type of problem. Has not fitted a comparison model or has

		comparison model for the problem that they were provided.	used a model for the wrong type of problem.
Model Evaluation	Use suitable methods to assess the performance of a model	Compared the performance of the two models/approaches using any method appropriate to the type of problem. Has described what the model comparison shows about the selected approaches.	Has selected a method not suitable for the type of problem. Has not described what the results show about the selected approaches.
Communication	Presents data concepts to small, diverse audiences	For each analysis step, has explained their findings and/or the reasoning for selecting approaches.	Has not provided a summary for each step (data quality, exploratory analysis, model fitting and model evaluation).

Sample Solution

You can find a sample solution from a published workspace link here. The sample solution demonstrates the required format for the final submission (i.e. a published workspace), and sufficient content needs to be included against the grading rubric. However, the sample solution is not the only solution.