sparkcognition

DATA SCIENCE CHALLENGE

Thank you for your interest in SparkCognition. The following DS Challenge allows our team to better understand your technical abilities and gives you some insight into the type of work SparkCognition's data scientists encounter on a daily basis.

You will have three (3) days to complete the exercise which is estimated to take approximately two (2) hours. Please aggregate your responses to a single notebook or file with associated comments and include your full name in the file name. Acceptable submission formats include .ipynb, .html, and .pdf.

Each section will be weighted equally in the evaluation. In addition, please note that you will be evaluated on your quality of coding, thought process, and clarity in communication.

Exercise 1: Data Preparation and Preprocessing

Please complete the following exercise, using comments in your code to explain your reasoning.

You have been provided with medical survey data that includes responses from an initial screening (demographic questionnaire), as well as a follow-up medical examination (blood pressure) with lab tests (cholesterol). This data is a subset taken from the NHANES 2005-2006 Survey and provided in the SAS XPT format. You can refer to the CDC website at https://wwwn.cdc.gov/Nchs/Nhanes/ and the links below for detailed descriptions of the features in each dataset.

Please clean and prepare the dataset for modeling, incorporating the following instructions.

Datasets:

Dataset	Link to files	Description
Demographics	DEMO_D.csv DEMO_RETIRED.csv	https://wwwn.cdc.gov/Nchs/Nhanes/2005-2006/DEMO_D.htm
Blood Pressure	BPX_D.csv	https://wwwn.cdc.gov/Nchs/Nhanes/2005-2006/BPX_D.htm

Cholesterol	TCHOL_D.csv	https://wwwn.cdc.gov/Nchs/Nhanes/2005-2006/TCHOL_D.htm

Instructions:

1. Use **python** to load DEMO_D.csv, BPX_D.csv, and TCHOL_D.csv as **pandas** data frames. Merge the 3 datasets but keep only the records that appear in all 3 datasets.
2. The existing features RIDAGEMN and RIDAGEEX contain the age of the respondent at screening and examination, respectively; however, both contain some missing values.
    a. Create a new feature AGE_AT_SCREENING with no missing values that contains the most precise estimate of each respondent's age (in months) at the time of screening. Choose an appropriate strategy to estimate the missing values **and** explain your reasoning.
    b. Create a second feature AGE_AT_EXAM with no missing values that contains the most precise estimate of each respondent's age (in months) at the time of the follow-up examination. Choose an appropriate strategy to estimate the missing values and explain your reasoning.
3. The existing features DMDEDUC3 and DMDEDUC2 contain categorical responses for each respondent's highest level of education completed: however, the categories are somewhat overlapping and complicated.
    a. Create a new categorical feature HIGHEST_EDUCATION with the following categories: ELEMENTARY (did not graduate HS, or currently in grades K-12), HIGHSCHOOL (graduated or GED), and COLLEGE (4 year graduates only). This feature should reflect the highest level of education **completed** for each respondent, from among the 3 options. Choose an appropriate strategy to fill in any missing values and explain your reasoning.
4. The file DEMO_RETIRED.CSV contains a single feature named RETIRED which is a binary flag indicating whether the respondent is retired (1) or not (0). This feature has some missing values.
    a. Suggest an appropriate strategy to fill in the missing values. Justify your approach using graphs or statistics.

sparkcognition

Exercise 2: Model Building

Please complete the following exercise in the same file as the other exercises, using comments in your code to explain your reasoning.

An auto-insurance company is revamping its pricing model. The analyst developing the new price model believes that the best approach is to develop 2 models: one for customers who are likely to file an insurance claim within the first year of their contract and another one for all other customers. The analyst has prepared a clean dataset consisting of 10,000 customers and 10 engineered features which capture driving behavior. The data has already been preprocessed for you (i.e., no missing data, no outliers, data is scaled, no correlated features, and the classes are fairly balanced).

The data is contained in *claim_prediction.csv*, where CLAIM = 1 means the customer filed a claim in the first year and CLAIM = 0 means the customer did not.

Develop a model to predict if customers will file a claim in their first year based on their driving behavior.

In addition to submitting your code, use comments to **explain** the decisions you made and how well you expect this model to perform on new data from a similar customer pool (and why).

Note that you are being evaluated on your model building and validation workflow, rather than on the complexity of your solution.

sparkcognition

Exercise 3: Model Evaluation

Please respond to the following short answer questions, numbered in the same file as other exercises.

1. What is one way to determine the number of clusters in K-Means clustering? How would you estimate the efficacy, or quality, of the K-means clustering results?

2. Your linear regression model is suffering from low bias and high variance. What steps can you take to improve your model?

3. Below is a scenario for training error (TE) and validation error (VE) for several iterations of a machine learning model. Which model would you choose, and why?

Model	TE	VE
1	105	90
2	200	85
3	250	96
4	105	85
5	300	100

4. You have built a model for a binary classification problem. The trained model was applied to the validation dataset and produced the results documented in the following confusion matrix.

n = 263		Predicted N	Predicted Y
Actual	N	97	48
Actual	Y	6	112

   a. Calculate Recall, Precision and F-1 score.
   b. If your classifier model is attempting to predict cancer in patients. Which type of error should you focus on for this type of problem? Which evaluation metric would you choose and why?
   c. If your classifier model is attempting to determine whether or not to recommend a YouTube video. Which type of error should you focus on for this type of problem? Which evaluation metric would you choose and why?

Exercise 4: Anomaly Detection

Please complete the following exercise in the same file as the other exercises, using comments in your code to explain your reasoning.

An oil and gas company with several offshore platforms is experimenting with anomaly detection on one of its platforms. An analyst has provided you with sample data for the pilot platform (*anomaly_detection.csv*). The data is a time series dataset, consisting of average daily readings from 5 sensors between 01/01/2016 and 12/30/2016. You can assume that data preprocessing (filling missing values, scaling, etc) has been handled by the analyst. You can also assume that the daily readings are independent and identically distributed.

The analyst has reviewed operation notes for the first 9 months (01/01/2016 to 09/30/2016) and identified that there were issues on the platform between 02/14/2016 and 02/21/2016. The analyst did not have time to review the final 3 months of data.

Use the first 9 months of data (01/01/2016 to 09/30/2016) to develop an anomaly detection model and test it on the final 3 months of data (10/01/2016 to 12/30/2016). How many anomalous periods were identified in the test period between 10/01/2016 and 12/30/2016? Only report anomalies that last longer than 2 days. Additionally, if an anomaly lasts for longer than 14 days, we consider that behavior to be the new normal, and we do not report it.

Hint: There are multiple potential approaches to solving this problem. One acceptable solution to this problem is to train a PCA model on normal data. Given new data, risk can be computed as the difference between the input data and reconstructed data, where reconstructed data is the result of compressing and decompressing the input data using PCA.