DOUGLASCOLLEGE

COMMERCE AND BUSINESS ADMINISTRATION

CSIS 3290-001: FUNDAMENTALS OF MACHINE LEARNING

Project #1: Data Wrangling and Regression Analysis

Due Date: Monday, June 15 by 6 pm

Learning Objectives

1. Extract data from online sources and prepare it into a form suitable for analysis
2. Apply regression analysis techniques to analyze the data and make predictions

Background

Given intense competition in the business environment, many companies utilize several strategies to stay competitive. Nearly every company now has an online presence. One the strategies companies use to win customers is through online search advertising. Using the search advertising strategy, companies often utilize two options for driving traffic to their website: (1) using **organic search** (i.e., using non-paid keywords), and (2) **paid search** or **Adwords keywords**(i.e., buying keywords from Google Adwords). Companies often spend hundreds of thousands of dollars on paid search, but the value they obtain from these strategies is not obvious.

The purpose of this project is to examine the business value of online search advertising strategies.

To do so, we are going to use Fortune1000 companies dataset since it provides outcome variables which can be applied to analyze the business value of online search advertising.

The project is divided into two parts. In part 1, you will collect data from online sources and prepare it into a form suitable for analysis. In part 2 of the project, you'll develop regression models, evaluate them, and make predictions with the chosen model.

This document provides a description of what you need to do for part 2 of the project. ***In this part of the project, you are going to examine whether online search strategies have impact on revenue and profits.***

Part 2: Regression Analysis and Making Predictions

In part 1 of this project, you extracted data from online sources and prepared the data into a form suitable for analysis. In this part 2, you are going to apply regression analysis techniques to analyze the data and make predictions with the selected model.

Instructions

- **Create a folder and rename it to include your name and course number (e.g., pAdams-CSIS3290-Prj1-Part2).**
- **All the files you are required to submit for the assignment should be placed inside this folder.**

Step 1: Read, Process, and Explore the Data

You are provided with a merged data file called "***df_merged.csv***". Please you this data file in your analysis. **Note**: If you completed part 1 of this project successfully, feel free to use your own data file. You must include it in the folder you were instructed to created above.

1. Read the data file into a DataFrame.
2. Please, select only the following columns of data to use in your analysis: Revenue($m), Profits($m), Assets($m), Market Value($m), Employees, Alexa Rank, Semrush Rank, Organic Keywords, Organic Traffic, Adwords Keywords, and Adwords Traffic.
3. Drop all rows with missing values from the data.
4. Process the data into numeric form suitable for analysis.
5. After you process the data into numeric form, check again if there are any missing values. If so, drop rows with missing values.
6. Compute "Average Rank" column of data by taking the average of the sum of Alexa Rank and Semrush Rank.
7. Explore the data by examining correlations among variables (please use seaborn heatmap), scatter plots, and histograms. Note that our target dependent (outcome) variables are Revenue and Profits. Focus on scatter plots between those outcome variables and particularly the following search-related variables: Organic Keywords, Organic Traffic, Adwords Keywords, and Adwords Traffic; and on the histograms of these 6 variables. **Note**: for the heatmap, control the size of the chart using matplotlib chart size options (hint: refer to heatmap in the udemy class example.)
8. Using a lambda function or other means, transform the data using mix-max normalization based on the following formula:
   Transformed data value = (x - min(x))/(max(x)-min(x))

Note: we are using min-max transformation because log transformation cannot be applied to negative and zero values. Also, min-max transformation does not fix skewness, it only helps with scaling of coefficients so that they are comparable.

Perform the rest of the analysis using transformed data from this step.

9. Using transformed data, estimate two explanatory models using the full scaled dataset (two models because we have 2 outcome variables: Revenue and Profits). For each model, use the following explanatory variables: Organic Traffic, AdWords Traffic, and Assets (proxy for company size). Why do you think we chose these variables only?
10. Split the data into training sample and testing sample, with 60% train and 40% test. Please use the "**random_state**" option to control the shuffling applied to the data before applying the split. Pass an integer value to this option for reproducible output across multiple function calls (common values used are 0 and 42. E.g., random_state=42). (see https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html#:~:text=train_test_split,-sklearn.model_selection.&text=Quick%20utility%20that%20wraps%20input,subsampling)%20data%20in%20a%20oneliner.&text=Allowed%20inputs%20are%20lists%2C%20numpy,sparse%20matrices%20or%20pandas%20dataframes).
11. Estimate OLS, Ridge and Lasso models using the train sample and assess their predictive performance using the test sample and RMSE statistic. **Note**: for Ridge and Lasso regressions, you will first need to determine the optimal regularization parameter (i.e., alpha) value to use). Again, you will need to estimate models using Revenue as the outcome variable and Profits as the outcome variable. Please use the same predictors as in 9) above.
12. From 11) above, select the best model and predict Revenue and Profits using the following input data:

Input	Asset ($m)	Organic Traffic	Adwords Traffic
Transformed value	0.25	0.15	0.21
Actual value	875830	142562520	1611306

Note: In your model, you use transformed values but in your Word document, you need to sue actual values, which people can understand. You will need to report actual predicted values, not their transformed counterparts (hint: re-arrange the mix-max formula to determine X and then use describe() method to get min and max values from the original DataFrame (i.e., the non-transformed DataFrame)

13. Prepare a Word document/report discussing the results from your analysis and their managerial implications. Make sure that you include relevant charts and tables from your analysis in your report.

What to submit (all in one folder, bearing your name as instructed above)

- Notebook/Python file
- CSV data file
- Word document