Brandon O'Briant
PREDICT 422-DL Practical Machine Learning
Assignment 1

Exploratory Data Analysis for Software and Course Preference

Introduction

Northwestern University's MSPA program consistently produces sought after graduates for data science roles by reviewing and updating course software choices and introducing new industry relevant courses to meet the current and future needs of the analytics industry. With the advent of the open source software, such as with the Python and R revolution reshaping the necessary skill set of the data scientist, this report utilizes results from the MSPA Software Survey collected in December 2016 to gain insight into potential new course and software offerings previous and current students would like to see offered in the program MSPA program.

Exploratory Data Analysis Methods

The MSPA Software Survey was collected in December 2016 and was conducted online using Survey Monkey. The survey dataset consists of 207 respondents with corresponding answers to questions (rows) and 40 columns representing survey questions. Removal of the outliers or transformations were not conducted since raw observations were originally related to a meaningful scale, but robust methods were included when looking at how one software or course preference correlates to another via scatterplots. Descriptive statistics were conducted for the main survey data: software preferences (a subset of survey data), potential new course offerings, and software preferences per graduate date (subset of software preference data). Skewness and kurtosis were utilized to gain an understanding of the distributions of non-categorical survey data. Boxplots of software and course offering preferences were constructed to gain an understanding of the distributions visually. Further, a correlation heat map, along with scatter plots, were constructed to evaluate which variable has a positive strong, positive weak, or

negative correlation when compared to the other software and course preferences. The top preferences were then checked against one another to see the percentage of the total preference they carry.

Overview of Programming Work

Python was used exclusively for the analysis of survey data, including the use of the following packages: Pandas and Numpy for data handling, while Matplotlib and Seaborn were used for visualization. The survey data was presented to us as a CSV file and loaded into the program using Pandas. The index was reset to correspond to respondent id's. Subsets of the data were taken via standard DataFrame slice operations, including .iloc. While, many built in functions in Python and the loaded packages were used to conduct many of the operations for data analysis and visualization purposes. Descriptive statistics were calculated via built in functions, while percentages were calculated via standard mathematical techniques with hardcoding. All the results were saved to separate .txt. files and visualization saved to separate pdf files.

Results and Recommendations

Looking at the overall preferences for software we see that Python and R showed the most preference, with SAS still hanging around as important. This is further backed when looking at these three software preferences per graduate year. Python and R preferences produced an overall highest percentage total for software preferences, with R still preferred above others; this could be due to the lack of familiarity with Python verses R. Then looking at which potential new course offering was rated the highest we see the Python Course as the top with 31.11%. Thus, my recommendation would be to offer more courses in Python, including Python programming foundations course.