2018FA_MSDS_422-DL_SEC58

Practical Machine Learning Assignment 1

Prabhat Thakur

1. Problem Definition

The MSPA program consist of 12 courses which uses different software and programming languages. It is important for program administrators to know which software students prefer more and there interest in potential new courses so that data science curriculum planning can be done to meet students interest as well as industry need.

2. Research Design and methods

Typically MSPA program courses use Java/Scala/Spark, Java/Script/HTML/CSS, Python, R, and SAS computer language or software systems. A survey was designed for students and faculty members to collect response for computer language or software systems in three categories: students desire to learn them, their professional need, and importance in industry. In each category, computer language or software systems are distributed in 100 points. Survey also captures student's interest in four new courses and their current program completion status. The data is stored in comma-delimited text file and consists responses from 207 students as of December 2016. Total student population eligible for survey is unknown.

The nature of this exercise is exploratory data analysis where collected data is analyzed to summarize their characteristics using statistical models and visual methods. Since we are not building predictive model, machine learning algorithm/models are not used in this exercise. We

analyzed the data from the three categories to understand which software or languages scored more in 100 point distribution and also which new course students like more.

3. Implementation and Programming:

For this analysis we used Python programming language and different python packages like numpy, matplotlib, pandas, and seaborn for data preparation and visualization.

The first step is to read the survey data from comma delimited txt file into python dataframe object followed by some data cleanup steps like converting respondent ids to row index, renaming column labels to short names for easy display in the plots and removing missing data. Scatter plots are generated using matplotlib package for personal preferences software and used seaborn package to examine inter-correlations among software preference in all three categories using correlation matrix/heat map visualization. Also calculated some statistics using pandas package describe() function. To see the software preference in all three categories and interest in new courses we also generated histograms to visualize point distribution from 0 to 100.

4. Findings and recommendations

From scatter plots of personal preference software and heat map we can see that Students who prefer Java also prefers Java scripts but least prefer R. Students who prefer Python least prefer SAS and Students who prefer R don't prefer other languages much.

Personal preference software histograms shows large number of students do not prefer Java and Java scripts languages. **The most favorite language is R, followed by Python and then SAS.** Similar trend can be seen in Professional need and Industrial importance categories. Students feel **R is most important language followed by Python and then SAS.**

Out of 4 new course, the most favorite course is **Python for Data Analysis** followed by **Foundations of Data Engineering.**