Using R for Analytics - MGMT 59000
R-Shiny App Final Project

Summer-2021

GitHub Link:
https://github.com/ryan-egbert/shiny_app_r
Video Presentation Link:
https://www.youtube.com/watch?v=c2F5gqk18QQ
Shinyapps.io Link:
https://regbert.shinyapps.io/college-major-recommender/

Submitted by -

Kai-Duan Chang (chang807@purdue.edu)
Ryan Egbert (regbert@purdue.edu)
Li-Ci Chaung (chuangl@purdue.ed)
Meghan Harris (harr1031@purdue.edu)
Gurijala Sri Manogna (sgurijal@purdue.edu)

Abstract
Each year, millions of high-school students face insurmountable pressure associated with choosing a college major. With a culture that follows a "Don't know it, Google it" philosophy, more and more high school students turn to outside resources to help them. If you google "What should my major be? "over 1.8 billion search results appear. Now add to this, the potential to have long-term financial, emotional, and economical consequences from this and it becomes an immobilizing decision. What if we could solve this problem for them? Our project tackles this increasing problem by providing a tangible and concrete solution to this question. By analyzing students' high school subject preferences, we recommend a major category for the student to focus on. We were able to do this through descriptive and predictive analytics from data that features various major categories such as salary and employment rate.

Business Problem
Every year, millions of high-school students need to decide what their college major should be. The problem is they often feel lost and have no direction to decide. The good news is that this problem can be solved through analytics. Our identified stakeholders are high-school students, college-students, and students looking to change their major. However, one constraint of this problem is that our project only focuses on favorite/least favorite subjects as our predictive variables.

Analytics Problem
Our analytics problem is to find a relationship between the major category and the high school subjects for the students; once this is established, we can recommend the college major based on their personal preference. We focused on three aspects to act as our drivers to connect to the output. These aspects are average salary, employment rate and the number of male/female students. The assumptions we held for this project include the notion that students want high salaries, care about employment rate, and consider class profiles before deciding a major. To assess the accuracy performance of our predictive model, we identified four different metrics to measure if our model is accurate and not over-fit. These metrics are "Hit ratio", "Accuracy Percentage, "AUC", and utilizing h2o.get_leaderboard/h2o.performance to assess the overall accuracy and competitiveness of the model amongst other models.

Data
We initially determined that our dataset should have a substantial number of variables and shouldn't require a significant amount of cleaning beforehand. We acquired our dataset, "College Majors" from kaggle.com, vetted it against our project requirements, and determined the data was sufficient to help us solve the business decision and analytics problem.

For data cleaning, we removed an irrelevant major category "Interdisciplinary Studies" which had no specific major within it and removed any rows that had missing values. Secondly, we split apart the major category "Computers and Mathematics" into two as we felt these specialties were distinct enough to stand as individual categories. Lastly, we identified, defined, and manually added unique values for each major category to use in our predictive model. These unique values were generalized high school subject categories for "Favorite Subject", "2nd Favorite Subject", and "Least Favorite Subject" columns.

Lastly, we used multiple caret functions to help us with our data transformation such as dummyVars(), findCorrelation(), and findLinearCombos(). From this, we found an interesting multicollinearity relationship between two variables. We found that "Arts" as a favorite subject and "PE" as a least favorite subject were highly correlated, so we removed this multicollinearity from our dataset.

Methodology Selection
For our project, we chose to focus on descriptive and predictive analytics. For descriptive analytics, our main goal was for the user to explore and compare different variables within each major category. We made it as easy as possible by providing effective data visualizations that allow for easy comparison. For predictive analytics, our goal was to recommend a major category for the student to focus on based on the user inputs. We, then,

provided a detailed analysis of all majors within the recommended major category based on salary and employment rate.

R was a viable tool to use for our project because it is a great interface to use for descriptive and predictive modeling. We utilized many helpful R packages and libraries, especially ggplot2 for our descriptive analytics and h2o for our predictive model. We chose to use these specific libraries because both libraries were easy to use and had seamless integration into our Shiny app.

Model Building

Because we only had a handful of predictors in our dataset, we wanted to ensure our model ran efficiently and accurately. To make our model simple and robust, we used h2o to create a Random Forest model. In our model we found that when we iterated through several values, we were able to find the best parameters (which were ntrees=50 and max-depth=5). We, then, integrated this model within our Shiny app to predict for our users. We also made sure the prediction would only populate until the user inputs all values and presses the action button.

Functionality

We have three tabs in our Shiny app that are "Explore", "Compare", and "Recommend". In our "Explore" tab, users can see the data overview, which includes detailed information of different majors, and the major/salary overview, which shows the major based on the salary range they chose. Users can also explore the distribution between salary and employment by major through a scatterplot graph. To make it very accessible, we also used the gghighlight and ggthemes packages to highlight only the chosen majors within the major category selection. In our "Compare" tab, users can compare their interested majors based on salary difference, employment rate, and male/female student difference. For our "Recommend" tab, users can pick their first, second, and least favorite high school subjects. Our predictive model, using the h2o package, then predicts the major category the student should focus their research on based on these inputs. To help them with this, we provided salary differences between the majors from the recommended major category. For future enhancements, we would love to include more data and data variables overall, but especially to add more predictive variables to our predictive model.

GUI Design and Functionality

Our Shiny app is simple, consistent and runs without errors. When the user needs to choose a major category, we used a drop-down list to avoid clutter. For salary, we used a slider so users can easily choose the range they want to view. For the scatter plot graph, we used a radio button so users can view all the categories and/or emphasize only the major category they want to compare. For our "Comparison" tab, we chose to use a bar chart to easily compare and see the difference between the selections. For our "Recommend" page, we kept the interface simple to help make the users feel comfortable when interacting with our app. We chose white as the main background color, so our colorful charts and tables stand out.

Conclusion

With our Shiny app, students have a guide that helps them choose a college major based on their actual interests. While we can't say for a fact that our Shiny app rivals the combined 1.8 billion google results from "What should my major be?", we firmly believe our project is a great starting place in tackling this challenging question.

References

We based our Shiny app on Zimin Luo's and Lasha Gochiashvili's project "Graduate Employment in Singapore" to build our UI from. We really thought their GUI design was perfect for how we wanted our interface to be for our project.

Link to Shiny App: https://gesurvey.shinyapps.io/Graduate-Employment-Survey/#1_brief_introduction
Link to GitHub Code: https://github.com/LashaGoch/R-Shiny-App-Graduate-Employment-Singapore