Project Proposal

Interesting classification problem. Don't forget that multiple classification algorithms needs to be used (3-4). And in the end you will need to choose your winning model according your evaluation criteria, and then make your interpretations of course. On your final project please emphasize a bit more why is this an important problem. Further, I don't see any literature review. Since it's an important problem, I'm sure others have worked on this problem. Make sure to also have this in your final project.

Logistic Regression

Decision Tree

K-Nearest Neighbors

<https://www.geeksforgeeks.org/logistic-regression-in-r-programming/>

<https://www.geeksforgeeks.org/k-nn-classifier-in-r-programming/>

<https://www.geeksforgeeks.org/decision-tree-classifiers-in-r-programming/?ref=rp>

<https://www.researchgate.net/publication/362280362_Salary_Prediction_in_Data_Science_Field_Using_Specialized_Skills_and_Job_Benefits_-A_Literature_Review>

<https://www.forbes.com/sites/cognitiveworld/2021/10/09/creating-curating-and-optimizing-with-data-science-and-machine-learning-at-the-new-york-times/?sh=5e069be52020>

Research Question:

Since the start of the coronavirus pandemic, how have different data-related full-time jobs are structured salary-wise on the tier of salary. Essentially, I am figuring out which data jobs are classified to which salary tiers to find their economic importance since 2020.

Data:

I will use the specific data called the Economics of Data Careers on Kaggle by Jonathan Bown but refine it down to only including data from the United States of America, fulltime jobs, and seven predictors. I will be investigating a classification problem.

Response variable:

The response variable will be what tier of salary. For the purpose of this research paper, we will use the breakdown from the U.S. News as a reference. We will consider the lower tier as incomes 0 – 80, 000 U.S. dollars. We will consider the lower middle tier as 80,000 – 200,000. The higher middle tier will be between 200,000 to 400,000. The high tier will be any salary greater than 400,000. Now we consider 100000 and above vs below

Predictors:

The 7 predictors will be the work year (2020, 2021, 2022), experience level (EN, MI, SE), job title, salary, employee residence (state), remote ratio (0, 50, 100), company size (small, medium, large).

Sample Size:

The sample size is about 300.

References:

<https://www.kaggle.com/code/jonbown/economics-of-data-careers/data>

**Final Project**

We will consider only papers which analyze economic related questions.

In the *References* section at the end of this document, you can find a paper and a book that have been written from economists and use machine learning techniques. This material and the references listed therein might be helpful in developing ideas about your project paper.)

A substantial portion of your grade ‐ 35% in total ‐ will be based on your final paper. The term paper should be **returned in R Markdown format** and the upload should 2 printable pages. The paper is due on **Monday December 5 on canvas. A research proposal is due on October 30.**

**Logistics of Submitting your project**Each student should submit her/his **own** copy. However, you need to mention **all the members**

of your group.

**Designing your project – Project Proposal (deadline October 30)**

At the stage of designing your project, you should think what your main question will be, what data you will use, and why this is interesting.

The project proposal should be uploaded on canvas by October 30. The proposal should be no more than a page outlining the question you will study, a description of the dataset you will use (sample size, response, predictors) with a link to it, and whether it is a regression problem or classification problem.

page1image5452592page1image5453216page1image5446768

**Evaluation**

This list is by no means exhaustive. A solid, coherent, creative paper will receive an A. Here are some tips on what is expected:

Abstract:

Since the start of the coronavirus pandemic, how have different data-related full-time U.S. jobs are structured salary-wise on the tier of salary based on remote-ratio, company size, position level, and other predictors is what I hope to find out. Essentially, I am figuring out which data jobs are classified to which salary tiers (below 100,000 U.S. dollars and greater than or equal to 100,000 U.S. dollars) to find their economic importance in the U.S. since the pandemic. This in terms of value on the job market as well as evaluate how company size, remote ratio, and skill level affect data job salary. I hope to better understand the growing demand for data jobs as students are wary of choosing such a field since they believe it is not financially secure. This is a question specific to the post-pandemic world that I am going to investigate as a classification problem. This will allow me to interpret data job trends and how the factors mentioned above affect potential salary and growth post-pandemic. I will use the specific data called the Economics of Data Careers on Kaggle by Jonathan Bown (<https://www.kaggle.com/code/jonbown/economics-of-data-careers/data>). I will refine it down to only including full-time jobs from the United States of America. The response variable will be the salary tier of either less than 100,000 (0) or greater than or equal to 100,000 (1). This is because students usually categorize a high salary by being six digits or more. For the purpose of this research paper, I will use the breakdown from the U.S. News as a reference. I will consider the lower tier as incomes 0 – 99, 000 U.S. dollars. I will consider the higher tier as 100,000 – infinity U.S. dollars. The nine predictors will be the work year (2020, 2021, 2022), experience level (EN, MI, SE, EX as dummy variables), remote ratio (0, 50, 100), company size (small, medium, large as dummy variables). The sample size is 328. I will conclude what the best model is out of a Logistic Regression Model, a K-NN model, and a Decision Tree Classifier model.

References:

Project Question:

Introduction:

My research question is centered on how valuable fulltime data are to companies since the pandemic in terms of salary depending on company size, remote ratio, and experience level. This will help me better understand the need for data-related skills for different companies and positions since the pandemic, where lot of data was needed to be processed. People should care if they are skeptical about whether a data job does not pay well or is not rising in value. People are also skeptical if remote ratio affects salary. In our research, we considered salary of 100,000 and above as high. This will hopefully instill more confidence in those wary of considering a data-related job because of job demand and salary. Machine learning is helpful in researching this question because I want to predict how remote ratio and company size can affect salary, which is why the data from 2020-2022 is useful.

Review of Literature:

What have other researchers said on the subject? You must include citations. Citations can be originated from popular news sources (such as *The Economist*, *The New York Times*, or *The Wall Street Journal*) and/or academic papers.

There has been another machine learning research paper called “Salary Prediction in Data Science Field Using Specialized Skills and Job Benefits” that aims to predict data science field salaries using specialized skills and job benefits as factors. They used random forests in order to have more stable and accurate prediction through the combination of multiple classifiers. Their focus also centered around data during and after the pandemic, which they call an industry revolution for how it changed and optimized jobs throughout the world to adapt to changing times. They claim there is a shortage of data science professionals and that through salary prediction for specifically data-related jobs they can allow students understand that those skills are valuable. The research paper essentially concluded that the best and most accurate classifying of salary actually depends on the type of data. They did not really interpret the data as they were more focused on finding the best model and deciding further testing is necessary.

Raheem, Mafas. *Salary Prediction in Data Science Field Using Specialized Skills and ...* July 2022, https://www.researchgate.net/publication/362280362\_Salary\_Prediction\_in\_Data\_Science\_Field\_Using\_Specialized\_Skills\_and\_Job\_Benefits\_-A\_Literature\_Review.

The news source *Forbes* has an article called “The Evolution of Data Science and AI at The New York Times” to highlight the increasing demand and necessity of data science jobs. They delve into how *The New York Times* has shifted to use artificial intelligence and data science in both the marketing and other business departments. This integration has only increased as businesses adapt to the availability of user data. They hold greater importance for data scientists because they want to know what data should be used to make marketing and business decisions. This includes analyzing risk and variability. They argue believe data science jobs are in demand and necessary for businesses in the future because of the value they provide.

Schmelzer, Ron. “The Evolution of Data Science and AI at the New York Times.” *Forbes*, Forbes Magazine, 21 Apr. 2022, https://www.forbes.com/sites/cognitiveworld/2021/10/09/creating-curating-and-optimizing-with-data-science-and-machine-learning-at-the-new-york-times/?sh=5e069be52020.

<https://www.researchgate.net/publication/362280362_Salary_Prediction_in_Data_Science_Field_Using_Specialized_Skills_and_Job_Benefits_-A_Literature_Review>

<https://www.forbes.com/sites/cognitiveworld/2021/10/09/creating-curating-and-optimizing-with-data-science-and-machine-learning-at-the-new-york-times/?sh=5e069be52020>

References

<https://www.researchgate.net/publication/362280362_Salary_Prediction_in_Data_Science_Field_Using_Specialized_Skills_and_Job_Benefits_-A_Literature_Review>

<https://www.forbes.com/sites/cognitiveworld/2021/10/09/creating-curating-and-optimizing-with-data-science-and-machine-learning-at-the-new-york-times/?sh=5e069be52020>

<https://www.kaggle.com/code/jonbown/economics-of-data-careers/data>

Bown, Jonathan. “Economics of Data Careers 💰.” *Kaggle*, Kaggle, 23 July 2022, https://www.kaggle.com/code/jonbown/economics-of-data-careers/data.

Raheem, Mafas. *Salary Prediction in Data Science Field Using Specialized Skills and ...* July 2022, https://www.researchgate.net/publication/362280362\_Salary\_Prediction\_in\_Data\_Science\_Field\_Using\_Specialized\_Skills\_and\_Job\_Benefits\_-A\_Literature\_Review.

Schmelzer, Ron. “The Evolution of Data Science and AI at the New York Times.” *Forbes*, Forbes Magazine, 21 Apr. 2022, https://www.forbes.com/sites/cognitiveworld/2021/10/09/creating-curating-and-optimizing-with-data-science-and-machine-learning-at-the-new-york-times/?sh=5e069be52020.

**Review of Literature:**

**Empirical Application:**

**For my classification problem, I will be using the classification algorithms of Logistic Regression, K-Nearest Neighbors, and Decision Tree Classification.**

Interesting classification problem. Don't forget that multiple classification algorithms needs to be used (3-4). And in the end you will need to choose your winning model according your evaluation criteria, and then make your interpretations of course. On your final project please emphasize a bit more why is this an important problem. Further, I don't see any literature review. Since it's an important problem, I'm sure others have worked on this problem. Make sure to also have this in your final project.

**Interpretation:**

**Empirical Application**: Much of this class is devoted to regression and classification with a primary view to obtain learning methods with improved prediction, for instance variables subset selection, regularized methods, tree‐based approaches. You should be able to apply these techniques, and others, using the data set you have chosen. To receive a good grade, these concepts must be applied *correctly*. In addition, the steps of analysis to reach the final best candidate model (or models) should be clear and coherent.

**Interpretation:** You will **not** receive a good grade if you just present a bunch of numbers and leave it to the reader to come up with their own conclusions. You need to thoroughly explain the results to the reader.

Experience level for entry and experience level for Mid are significant in relation to salary. This means there is potential room for growth in the data science industry.