Group 26 Project Proposal

Authors: Jasmine Chen, Vedant Goyal, Usman Moazzam, Kanisha Shah, Lin Zhu

Background

In recent years, roles related to data science, including data scientists, data analysts, machine learning engineers, data engineers, and more, have seen remarkable growth and garnered significant attention for their pivotal roles in reshaping industries. Situated at the intersection of computer science, mathematics, and domain expertise, these roles have collectively become instrumental in extracting insights from vast datasets and driving data-driven decision-making. The descriptor "the sexiest jobs of the 21st century" gained prominence, notably in the article "Data Scientist: The Sexiest Job of the 21st Century," which was published in the Harvard Business Review in 2012[1] (Davenport & Patil, 2012). This phrase captures the allure and high demand associated with roles that encompass data analysis, machine learning, and data engineering. As data generation continues to surge due to technological advancements, these professionals have become indispensable, sought-after experts.

Understanding salary trends and factors influencing compensation is pivotal for both aspiring and current professionals in these fields. Our group project addresses this need by utilizing a comprehensive dataset that encompasses demographic information, technical skills, learning resources, and more. We aim to develop a machine learning model capable of predicting salaries for professionals in these data-centric roles. This initiative empowers individuals to make informed career decisions and assists organizations in attracting and retaining top talent in the competitive landscape of data-related professions, contributing to the continued growth and success of these roles in the 21st century.

Description of Data Set

The data came from a survey run by Kaggle in 2020. The survey was open for approximately one month and collected over 20,000 responses. In addition to collecting more basic information such as compensation, job title and size of the company, the survey includes data on the type of technology used by the professionals and their companies. This provides us with a rich dimension of information that we will be able to build model features on. In the survey, there are no user-entry responses, which allows for a simpler response parsing process. Single response multiple choice questions are recorded in a single column, and multi-select questions are broken down such that a particular column would be populated if the user has chosen that option. This translates very easily into one-hot encoding since non-empty columns can be marked as 1.

Proposed Machine Learning Techniques

Our endeavor would be to adopt a systematic approach in the exploration of diverse machine learning techniques. The methodological journey will commence with the application of linear regression, followed by support vector machines, decision trees, and various ensemble methods, encompassing random forests, bagging, and boosting. Subsequently, neural networks will be considered for analysis. A preliminary examination of the dataset permits the formulation of a prefatory hypothesis regarding the suitability of specific techniques. It is worth noting that linear, ElasticNet, Ridge regression etc. may prove suboptimal for this particular problem due to the predominantly categorical nature of the features. However, all aforementioned techniques will be subjected to scrutiny, given the relatively modest dataset size and manageable computational requirements. Essential to the selection of an appropriate machine learning technique is the careful consideration of evaluation metrics, a topic to be expounded upon in greater detail following the completion of preprocessing and visualization procedures.

---

[1] Davenport, T. H., & Patil, D. J. (2012, October). Data Scientist: The Sexiest Job of the 21st Century. Harvard Business Review. https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century