CSCI 5502: Final Project Key Results

Aaptha Boggaram, Arindrajit Paul, Nikola Tanovic

December 2023

1 Key Results

Generally, a score of 80% and above for all the performance metrics mentioned denote a model's good performance. Out of the 6 models that were used, the ensemble models, that is, RandomForest Classifier and Gradient Boosting Classifier, performed the best. This is to be expected because the final predictions of such models are the result of aggregation from multiple other models, hence the name ensemble. However, this robustness of the ensemble models should not imply that such models be used with skewed data, which is the case for our dataset.

In reality, skewed data can induce biased learning as this can misrepresent the true distribution of the data, which decreases the models' performance (as evident by the performance metrics). Notably, for the case of k-Neighbors Classifier, such performance is unexpected given that the algorithm is highly sensitive to skewed data. However, the general trend that we saw is that, as the value of k (number of neighbors) increases, the accuracy and the subsequent metrics go up as well, until around $k = 15$, where the numbers plateau. In conclusion, this dataset is more suitable for data trend analysis than predictive modelling, but if the latter is to be conducted, more preprocessing and model tuning is necessary to achieve better performance. Crucially, a greater volume of data is required to ensure ample resources for preprocessing, while simultaneously retaining sufficient data for predictive analysis.

In our analysis, we took in consideration the fact that $1 in the 1990s has the same purchase power of $2.30 today. This means $50k in the 1990s is roughly $115k in today's currency. In the data, significantly more people earn less than $50k a year, which is true for the case of US census data from 2021-2022. More specifically, we have made the following observations:

- The level of income is directly proportional to the education level of an individual. We saw an increase in the number of people with a bachelor's degree or higher earn more than $50k a year, suggesting that a college degree is a significant contributor in the level of income of an individual. This trend, however, is not consistent with other education levels, such as high school diploma or some college degree.

- A significant proportion of individuals younger than 35 years earned less than $50k a year and individuals older than 45 years earned more than $50k a year. This trend is similar for the census data of 2021-2022. Additionally, in both cases, the proportion of individuals in higher level occupational positions, such as executive managers and professional specialist, earning more than $50k were higher compared to other occupations.

- More than 90% of the individuals in the 1990s census data were native to US. Due to this, they had more opportunities at jobs that earned them more than $50k a year, in comparison to the immigrants who were not born in the US. However, the current census data shows that both the natives and the immigrants have had equal opportunities at jobs that pay more than $115k (more than $50k) a year.