EXECUTIVE SUMMARY

Random Forest on IT Salary Dataset

Data Analytic Graduate Capstone – TCM2

Carl Andrew Perkins

Western Governors University

C772

EXECUTIVE SUMMARY

Random Forest on IT Salary Dataset

A: Research Problem Statement and Hypothesis

According to Donges (2020), a Random Forest model is a supervised learning algorithm which merges decision trees in an ensemble type of methodology. Donges (2020) also found that utilizing a Random Forest regression model can help predict both classification and continuous variable alike. Aasim (2019) hypothesized that salary positions data can be descriptive enough to predict salaries while utilizing Random Forest. This study seeks to find whether or not a Random Forest model can predict salary of a Data Professional survey dataset.

The hypothesis is that a Random Forest model can indeed predict the salary of this dataset using pay band categories as a classification target variable.

B: Tools and Limitations

Python was used for the importing of and cleaning the data. Then Python created the Random Forest model. According to Kunal (2017), Python also has by far the most deep learning support. It is for this reason that Python has been selected as the tool of choice.

One limitation of this dataset is that this study is a collection of only four years of survey data. Another limitation of this study is that values are missing for certain fields. Several variables with the values “Not asked” were removed during this analysis.

C: Data Collection and Context

The data for this project is publicly-available information provided by Brent Ozar Unlimited and is a survey of Data Professional’s salaries dating back to 2017 (Ozar, 2020). Before removing outlying records, the dataset contains 8,628 rows.

Brent Ozar Unlimited has made this data publicly available as a part of public domain. There is no private information of any kind within the dataset that restricts use or identifies specific individuals. This data does contain some extreme outliers, particularly as it pertains to lower-bound and upper-bound salary outliers. These outlier have been removed during initial cleaning; other variable outliers have been handled by imputation. According to Grace-Martin (2020), there is sufficient evidence to have removed extreme outliers.

D: Data Preparation and Analysis

The Excel file was downloaded from the brentozar.com website. While a portion of outliers were removed, Kho (2020) suggest that this model in particular is not heavily influenced by them and, subsequently, was a good fit for this dataset. After the outliers were removed the data was, generally speaking, in good shape. The data mainly contains categorical variables but has a few discrete number values associated. Python has been used to clean the data, remove applicable variables, remove outliers, and generally scrub the data.

The most important analysis was of the salary variable itself. There were several extreme outliers that were removed before the rest of the variables were analyzed. A large distribution of missing “Not asked” were found for a variety of variables; some of these variables were removed completely from analysis. Due to different questions asked during different years, the data sparsity prior analysis and preparation was 24%. Indicating that prior preparation, the density or usable data was roughly 75%.

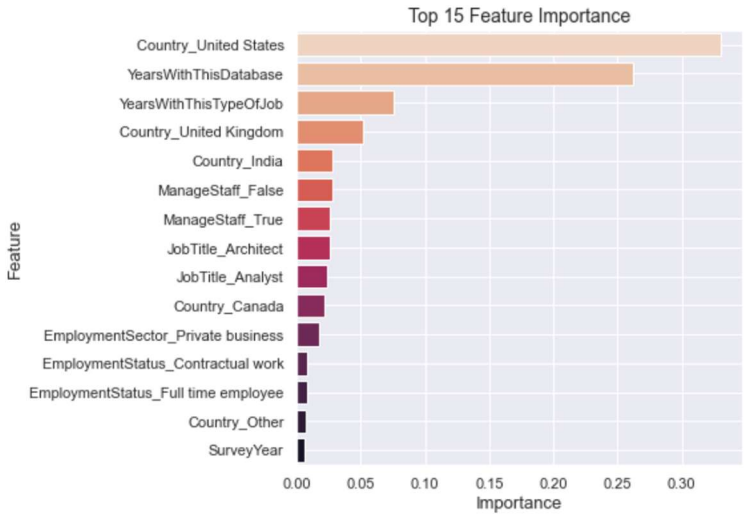
Every variable was methodically analyzed. Bivariate, Univariate, and Multivariate variable analyses, as well as normality tests, were conducted on each variable within the dataset. Each variable was inspected for outliers, as well as missing values for machine learning

accuracy. Some variables were also updated to specific and consistent data types across the dataset. After analysis and preparation, the sparsity was 6%, indicating that 95% of the scrubbed data was usable and helpful.

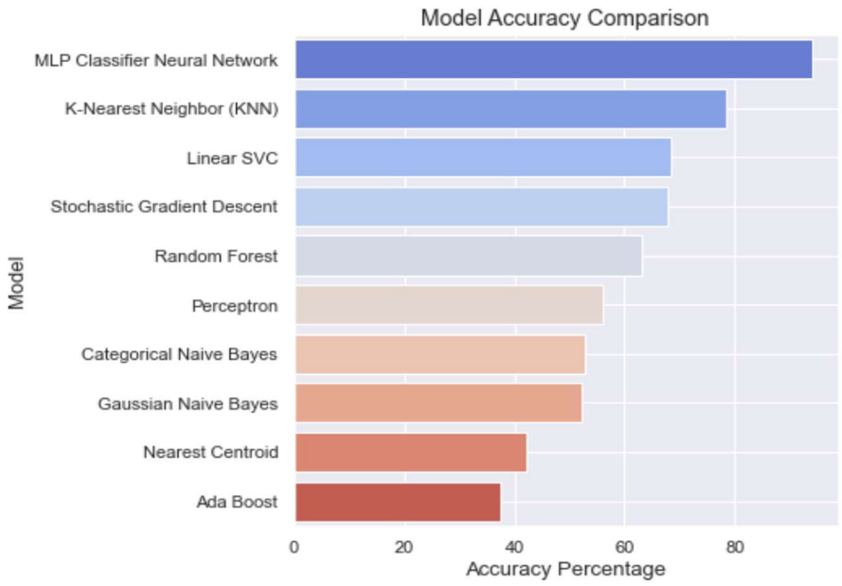
Principal Component Analysis (PCA) and Cluster Analysis were also used as unsupervised techniques on various variables. PCA was used to reduce the variables to those that explained the most variance. Cluster Analysis was used in conjunction with creating the Salary pay bands. Essentially the cluster analysis helped to discern how many pay bands (tiers) would be optimal within the data set as the target categorical variable. Lastly, nine other machine learning models were used to compare against the accuracy of the Random Forest learning model itself.

E: Findings

After splitting into five salary categories, the Random Forest machine learning model predicts with an accuracy of 63%, thereby confirming the original hypothesis. Additionally, the top three variables which had the most impact on the Random Forest selection process were United States and experience variables.



Lastly, while the Random Forest model does in fact predict salary, there are other models that are more accurate.



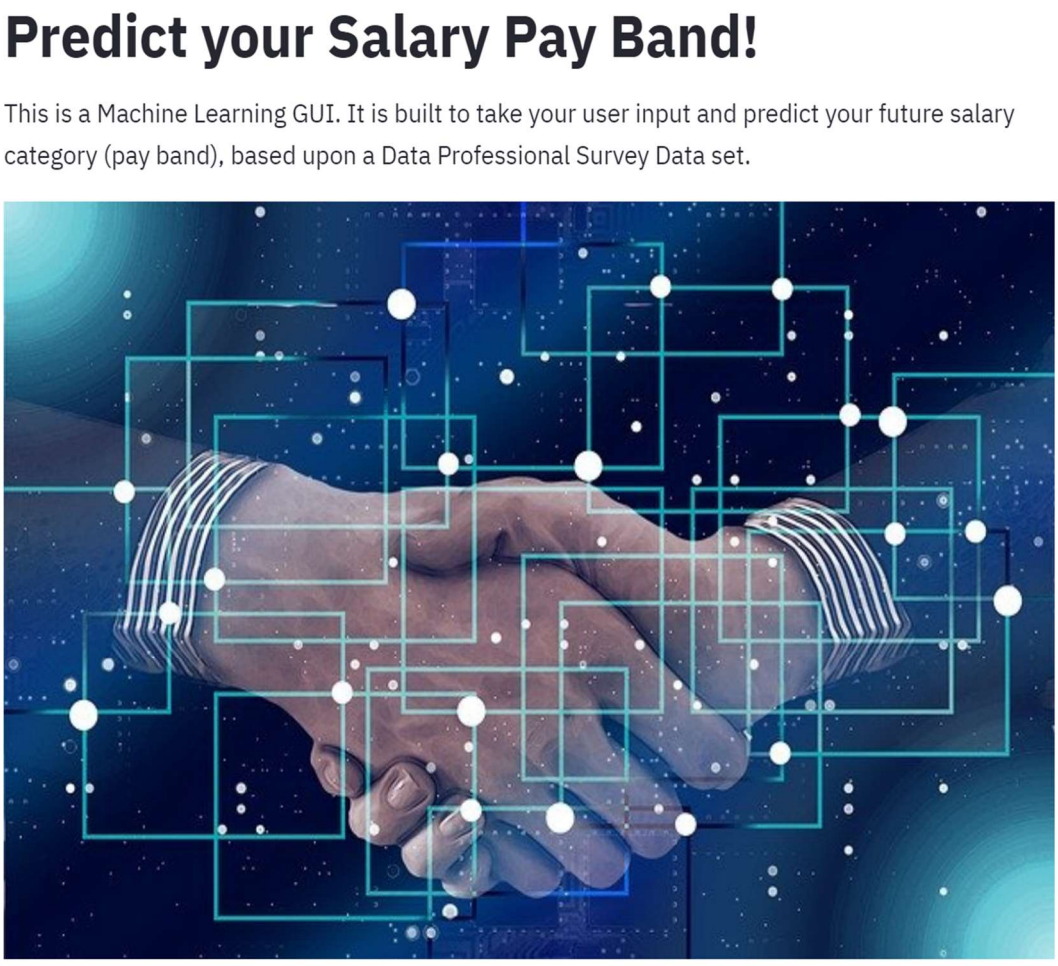
E: Proposed Actions

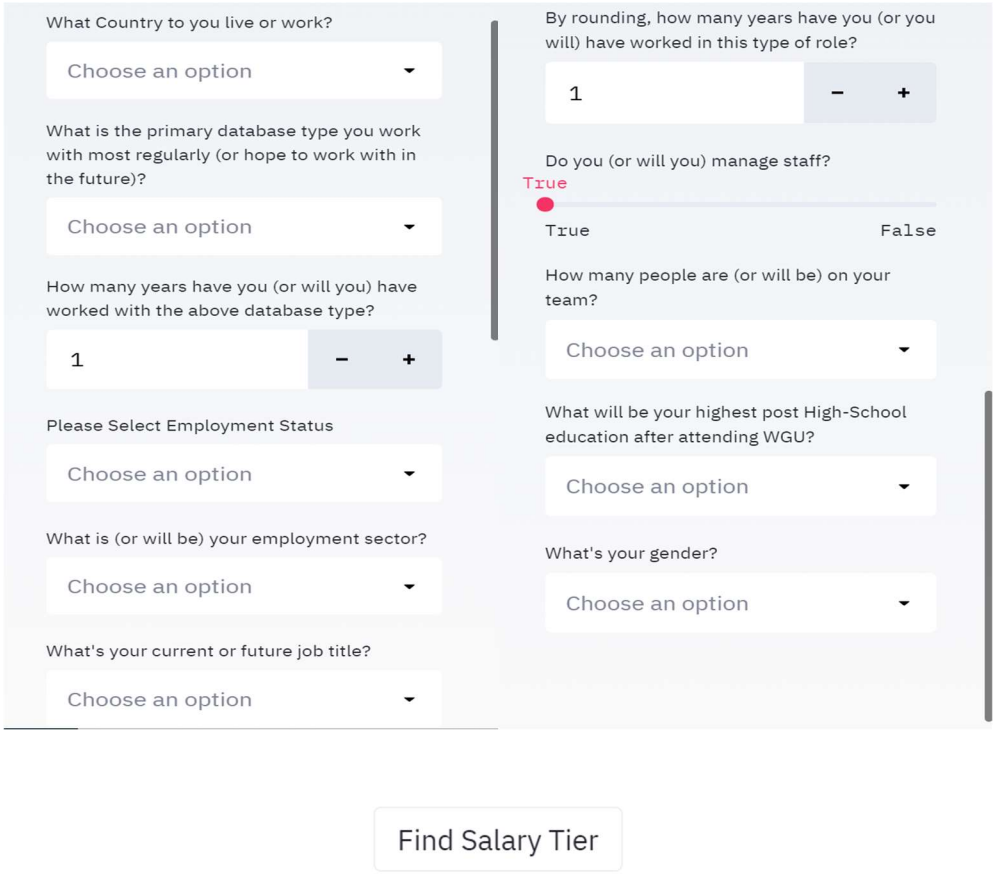
One of the proposed actions is to use a larger dataset. The accuracy of the model would likely improve if another larger dataset were used. Another option would be simply waiting on an accumulation of more years of the survey data. A reason for using another dataset all together is that survey data in general is prone to inaccuracies. By using data that isn't a survey, the underlying dataset would likely be more accurate overall.

A second proposed action would be to use one of the more accurate models as a prediction methodology. K-Nearest Neighbor (KNN) and MLP neural network would be the best two options of the machine learning models compared against. While the neural network is by far the most accurate, it is also very computational expensive. It is for this reason that the KNN is the best model to be used for a better prediction methodology.

F: Data Professional Salary Prediction Tool

Typically, when organizations move a machine learning model to production, the models' application isn't used particularly well. The creation of this second tool allows students to predict their own pay band after graduation, while also showing them real world application of a machine learning model. The program itself uses the streamlit API as a GUI in order to provide WGU students a way to predict their Salary Tier in future years. This is an implementation of the second proposed action, mentioned above. KNN was chosen for its speed and accuracy. This tool provides WGU and its students real value in predicting their salary while also showing a finished capstone analytic product. Below are screenshots of the tool itself.





Accurately answer the inputs in the web UI. Try and base your answers at the time of degree completion or two years from today. Then press the button above to initiate the machine learning model prediction. Note: Make sure to input all value variables in order to get a prediction. Additionally, erroneous values will cause an error (aka 100 years, etc.).

G: Summary of Benefits

This study and tool provide WGU and students real value. The KNN prediction model increases accuracy of prediction by 15+ %. The tool allows students to predict their future salary pay band, post-graduation. Lastly, it enables synergy by allowing Data Analytic Masters students at WGU to experience hands on an example of a machine learning tool carried through to production.

References

Donges, N. (2020, September 03). A Complete Guide to the Random Forest Algorithm. Retrieved Janurary 05, 2021, from https://builtin.com/data-science/random-forest-algorithm

Aasim, O. (2019, September 06). Machine Learning Project 6: Predict Salary using Random Forest Regression. Retrieved Janurary 05, 2021, from https://medium.com/analytics-vidhya/machine-learning-project-6-predict-salary-using-random-forest-regression-9a18f97c91e5

Kunal, J. (2017, September 12). Python vs. R vs. SAS – which tool should I learn for Data Science? Retrieved Janurary 05, 2021, from https://www.analyticsvidhya.com/blog/2017/09/sas-vs-vs-python-tool-learn/

Ozar, B. (2020, January 05). The 2020 Data Professional Salary Survey Results Are In. Retrieved November 03, 2020, from https://www.brentozar.com/archive/2020/01/the-2020-data-professional-salary-survey-results-are-in/

Grace-Martin, K. (2020, November 03). Outliers: To Drop or Not to Drop. Retrieved Janurary 05, 2021 from https://www.theanalysisfactor.com/outliers-to-drop-or-not-to-drop/

Kho, J. (2018, October 19). Why Random Forest is My Favorite Machine Learning Model. Retrieved Janurary 05, 2021, from https://towardsdatascience.com/why-random-forest-is-my-favorite-machine-learning-model-b97651fa3706