Comparison of Linear Regression and Decision Tree to predict Salaries

Nidhi Joshi (MSc Data Science - FT)

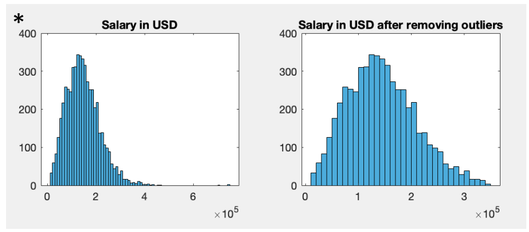
Description and Motivation

Solving and quantifying the relationship between various factors related to jobs and resulting salaries through regression analysis models like Linear Regression and Decision Tree models. Analysing each model's performance through performance metrics and comparing the results with each other. Reviewing and comparing results from a study by [multiple authors][1] on a similar dataset. This analysis was done with an aim of providing guidance to the professionals wanting to pursue the field for career progression.

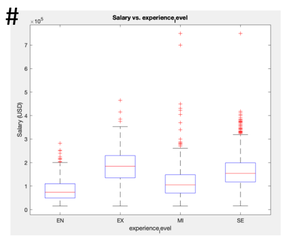
Exploratory Analysis

- Dataset is taken from Data jobs salaries (weekly updated) from Kaggle[2]
- This dataset originally retrieved from https://ai-jobs.net/ has anonymous salary information of Data Science professionals around the world
- Although, since large real world salary datasets are quite scarce, there is very little or no research about predicting salaries[3]
- It originally contained 9556 rows with 10 features and 1 target column. 4 columns were numeric in nature including the target column, and 7 were categorical including one column(company_size) which was ordinal in nature
- The data was cleaned to remove the null values and duplicate rows shrinking the row size to 5488
- The target column is respective salaries converted to one currency (USD) for the sake of better comparison
- Feature selection was done by box plotting and scatter plotting the respective variables against the target - Salary (in USD)
- Job title and Salary (in home currency) were dropped as no direct relation was established with the target column
- The histogram indicates that the target variable (salary_in_usd) is right skewed but well distributed otherwise as shown in the figure*
- Some plots indicate a strong relationship between the salaries and some related features. It can be analysed through the displayed box plots and the scatter plots here#
- The following deductions can be made through the plots:
  - Medium size' companies are paying the most on an average
  - People at 'Executive' position are paid the most among all categories
  - 'Full-time' employees are earning well as compared to others
  - Location-wise, Puerto Rico and US for both employee and employer locations are good payers but these two features aren't displaying a distinct relation with the salary
  - Those paid in 'USD' currency have a par above those paid in other currencies
  - While those working either remotely or offline in full capacity are being paid almost the same with Offline employees having a little edge over remote workers; but those working in hybrid models are being paid lesser
  - With increasing years from 2020 to 2023, the salaries have increased too

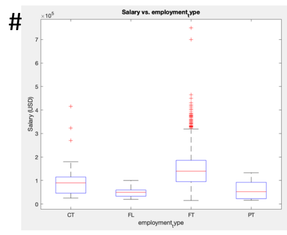
*



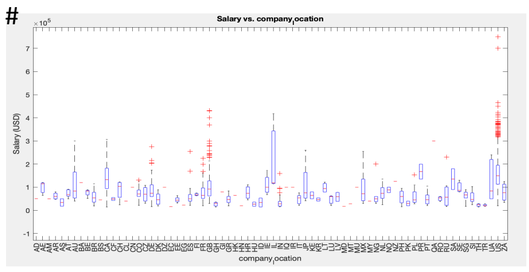
#



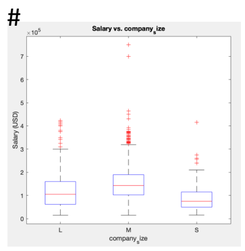
#



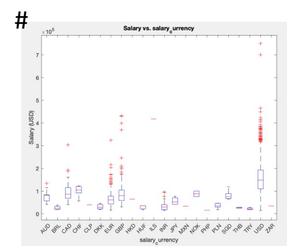
#



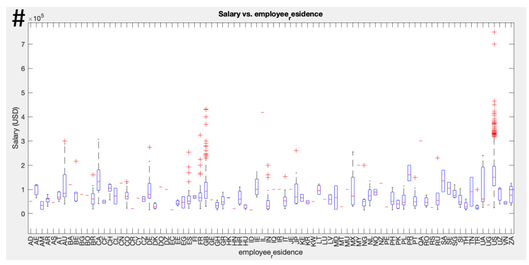
#



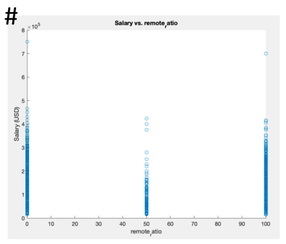
#



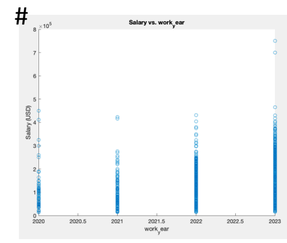
#



#



#



Linear Regression

- Linear Regression is used as a training model in case of supervised machine learning
- It starts with analysing the correlation between independent factors (features) and dependent output (target)
- Then, this algorithm models a linear relationship between all the variables and the target by a statistical linear equation - $y_i=\beta 0+\beta 1x$, where x = independent variable, y = target variable, $\beta 0$ = slope and $\beta 1$ = intercept[4]
- Post tuning the model to fit our data better, the efficiency of the model is finally evaluated through R-squared and RMSE parameters

Pros

- ⊕ One of the simplest models to implement and hence quick to train
- ⊕ Since no complicated calculations are involved, hence lesser modeling speed even when the data size is huge
- ⊕ With features being independent of each other, it is easy to comprehend the changes in the target with any change in the predictor variables

Cons

- ⊖ The real-world data is much more complex for a simple model like Linear Regression[5]
- ⊖ Assuming linearity and homoscedasticity between feature and target variables may not be realistic in many cases
- ⊖ This model is not capable of minimising the effects of outliers in the calculations, hence the results are not reflective of real information[6]

Decision Tree

- Decision tree is a hierarchical model type of supervised machine learning
- Made of nodes and branches where nodes are the questions and branches are the answers to them
- In a top-down approach, Decision tree model splits the data in a recursive manner until most of the information is divided under different class labels[7]
- Pruning method can be used if overfitting is an issue and, to bring down the model complexity. Pruning helps eliminate unimportant branches

Pros

- ⊕ Very flexible in handling both categorical and numerical data
- ⊕ Very little or no data pre-processing required
- ⊕ Decision tree visualizations are easier to understand as compared to other models
- ⊕ Capable of handling complex relationships and variables dependencies

Cons

- ⊖ With large datasets, the calculations can lead to very large trees hence, increasing the cost and storage requirements
- ⊖ The results can be misleading if the data is skewed
- ⊖ Even with slight changes in data, the output can be very different
- ⊖ Evaluating the performance of the model can be difficult in case of too many trees

Hypothesis Statement

- Decision Tree to spend slightly longer time to train as compared to Linear Regression
- Decision Tree is expected to perform better owing to its ability to handle both categorical and numerical data efficiently
- The results for both models may be almost similar as the mathematical formula for decision tree regression is like the linear regression model. Standard deviation and variance calculations are used to define the decision tree nodes[8]
- Assuming the residual is minimum, the linear regression model would aim to find the best coefficients for the variables to predict the closest outcome possible[9]

Methodology

- After pre-processing the data, split the given dataset randomly in 80:20 ratio as train and test data respectively
- This divides the data into 4391 rows to train and 1097 to test after building the model
- Cross-validate the training data using Holdout as the dataset is relatively smaller
- Select relevant features and identify machine learning models along with the required hyperparameters to better optimise the models
- Evaluate both the models through performance metrics like R-Squared and Root Mean Squared Error
- Predict the results on the test set and visualise the actual vs. predicted values along with the residuals
- Compare both the models with respect to the performance metrics, training time and residual error

Choice of parameters and experimental results

➢ Linear Regression

- Feature scaling done on the target variable as there were a lot of outliers and categorical features converted
- Fitting the regression model by Ridge regularization technique (L2 regularization) with cross validation on the training set
- Assessing the model performance through metrics like RMSE and R-Squared

Choice of parameters:

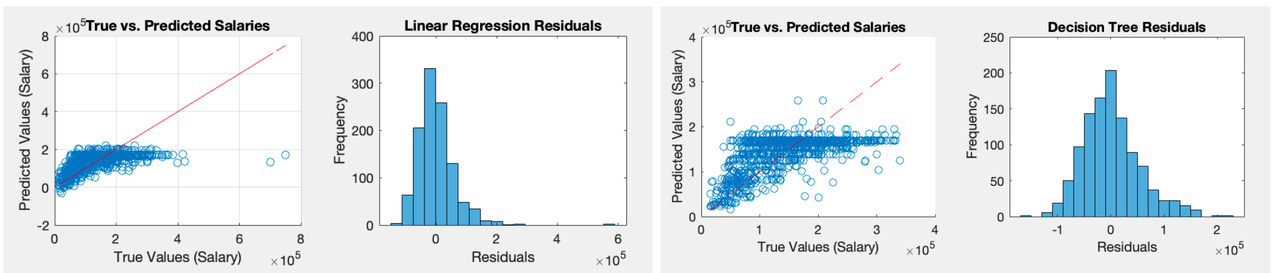
- Regularization – Ridge
- Cross-validation by KFold – 5

➢ Decision Tree

- Fitting the Decision Tree model by passing predictor features and response variable.
- Iterating a test model to find out the best leaf size through a defined set of values and comparing with each MSE value.
- Pruning the decision tree to a reduced depth to get better performance

Choice of parameters:

- Leaf size – 5
- Pruning depth – 5



Analysis and evaluation of results

- In terms of RMSE, the model performance for Decision Tree is almost 13% better than the Linear Regression model. This can be owed to the fact that Decision tree is efficient in assessing which variables are of better importance to predict better. With the random sample of training data which is cross-validated too, any bias in calculations is avoided by tackling imbalanced data as well
- The results as mentioned in the table infer that Decision Tree model was the most effective fit to the data, with a greater R-squared value and lesser RMSE and MAE
- Since most of the features were categorical, one-hot encoding needed to be done for the Linear Regression model which changed the number of features to 194. One hot encoding formats the data in a way that it eases the model training process. "It produces a binary vector of categorical variables with a value of 1 for each row having this option and 0 otherwise."[10] However, Decision Tree model was capable of handling both the categorical and numerical values and hence, no such treatment was required. But the test datasets become different for both now
- Furthermore, to better predict through the models, only relevant features were selected. As we can see from the box plots, most features are equally important to predict the salaries, hence, Ridge regularization was used in linear regression to avoid any feature exclusions
- There were several outliers detected when the salary was plotted which were weeded out to optimise the models. The performance of the models slightly improved after the optimisation
- The best leaf size for the Decision tree was determined by comparing KFoldLoss for a pre-defined set of values. Further pruning the tree to a depth level of 5 helped improving the average performance of the model. To avoid the over-fitting issue that Decision tree model often faces with larger tree sizes, smaller trees are preferred, "which is consistent with the principle of parsimony in Occam's Razor; that is, "entities should not be multiplied beyond necessity.""[7]
- While predicting values on the test set through the final models, surprisingly Linear Regression took more training time as compared to the Decision tree model. This may be due to the increased number of feature columns as required for carrying out Linear Regression with categorical values in the dataset
- Comparing the final model predictions, the residual errors for Decision Tree are considerably lower than those in Linear Regression even though none of them look like an ideal prediction for this dataset when we look at the Actual vs predicted values plot. But if we compare both, Decision Tree is a better alternative for this dataset
- A study[11] demonstrated that *R*-squared (or coefficient of determination) provides with more information and is a true metric as compared to MSE, RMSE, MAE and MAPE which have interpretability limitations. So, even if we only use R-squared as the standard metric, Decision tree performs 15.3% better than our counterpart Linear Regression model
- Due to the simplicity of our final models and reduced number of features, they suffered from high bias and low variance which means that they don't fit the training and testing data very well.

Model	R-Squared value	RMSE value	MAE value	Time taken (one-time)
Linear Regression	0.2965	60818	43088	127210
Decision Tree	0.3418	52895	40663	124590

Lessons learned and Future work

➢ Lessons

- When it comes to handle categorical data where the features in question are a lot more than expected, Linear Regression is not a very good model to consider for prediction. It is better to look at models like Random Forests instead
- Correlation matrix could have had provided with a better comparison in variable dependencies in the data for the purpose of feature selection

➢ Future

- Acquiring a larger dataset would be very useful in training better performing models giving way to training and validating a larger scope of information[12]
- Feature selection to be employed with methods like Correlation matrix for some robust results
- Gradient boosting can be explored on decision trees algorithm to avoid any separate encoding or pre-processing on the categorical features

References

[1]N. Niknejad, M. Kianiani, N. P. Puthiyapurayil and T. A. Khan, "Analyzing Data Professional Salaries Exploring Trends and Predictive Insights," 2023 International Conference on Big Data, Knowledge and Control Systems Engineering (BdKCSE), Sofia, Bulgaria, 2023, pp. 1-6, doi: 10.1109/BdKCSE59280.2023.10339759.

[2]VÁZQUEZ, L. (2023) *Data jobs salaries - weekly updated, Kaggle*. Available at: https://www.kaggle.com/datasets/lorenzovzquez/data-jobs-salaries (Accessed: 19 December 2023).

[3]Eichinger, F. and Mayer, M. (1970) *Predicting salaries with random-forest regression*, *SpringerLink*. Available at: https://link.springer.com/chapter/10.1007/978-3-031-18483-3_1 (Accessed: 19 December 2023).

[4]G. Wang, "Employee Salaries Analysis and Prediction with Machine Learning," 2022 International Conference on Machine Learning and Intelligent Systems Engineering (MLISE), Guangzhou, China, 2022, pp. 373-378, doi: 10.1109/MLISE57402.2022.00081.

[5]*What is Linear Regression?* (1978) *Amazon*. Available at: https://aws.amazon.com/what-is/linear-regression/#:~:text=Linear%20regression%20is%20a%20data,variable%20as%20a%20linear%20equation (Accessed: 19 December 2023).

[6]Satyavishnumolakala (2020) *Linear regression -Pros & Cons*, *Medium*. Available at: https://medium.com/@satyavishnumolakala/linear-regression-pros-cons-62085314aef0 (Accessed: 19 December 2023).

[7]*What is a decision tree* (no date) *IBM*. Available at: https://www.ibm.com/topics/decision-trees#:~:text=Decision%20tree%20learning%20employs%20a,classified%20under%20specific%20class%20labels (Accessed: 19 December 2023).

[8]N. Niknejad, M. Kianiani, N. P. Puthiyapurayil and T. A. Khan, "Analyzing Data Professional Salaries Exploring Trends and Predictive Insights," 2023 International Conference on Big Data, Knowledge and Control Systems Engineering (BdKCSE), Sofia, Bulgaria, 2023, pp. 1-6, doi: 10.1109/BdKCSE59280.2023.10339759.

[9]A. Jain, S. Jain, N. M. Pancinovia and J. P. George, "A Non-Linear Approach to Predict the Salary of NBA Athletes using Machine Learning Technique," 2022 International Conference on Trends in Quantum Computing and Emerging Business Technologies (TQCEBT), Pune, India, 2022, pp. 1-5, doi: 10.1109/TQCEBT54229.2022.10041664.

[10]R. Kablaoui and A. Salman, "Machine Learning Models for Salary Prediction Dataset using Python," 2022 International Conference on Electrical and Computing Technologies and Applications (ICECTA), Ras Al Khaimah, United Arab Emirates, 2022, pp. 143-147, doi: 10.1109/ICECTA57148.2022.9990316.

[11]Chicco, D., Warrens, M.J. and Jurman, G. (2021) *The coefficient of determination R-squared is more informative than smape, Mae, MAPE, MSE and RMSE in regression analysis evaluation*, *PeerJ. Computer science*. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8279135 (Accessed: 19 December 2023).

[12]N. Niknejad, M. Kianiani, N. P. Puthiyapurayil and T. A. Khan, "Analyzing Data Professional Salaries Exploring Trends and Predictive Insights," 2023 International Conference on Big Data, Knowledge and Control Systems Engineering (BdKCSE), Sofia, Bulgaria, 2023, pp. 1-6, doi: 10.1109/BdKCSE59280.2023.10339759.