Comparing Different Machine Learning Models to Determine Best Model for Predicting Salary

Ruohan Huang, Jacob Nguyen, John Sebastian Solon,
Frank Wem Guang Zhu, Amritpal Zenda

December 10, 2024

1 Abstract

This study investigates salary prediction using machine learning models based on demographic information and professional attributes. We used a dataset of 6,684 entries containing information about age, education, experience, job titles, and other factors, we developed and compared multiple prediction models. After addressing dimensionality challenges through feature engineering and PCA, we implemented and evaluated four models: Linear Regression, Polynomial Regression, Neural Networks, and Random Forest. Our analysis revealed that Random Forest achieved the best performance with an R^2 score of 0.848 and MSE of 422,546,711.45, followed by Neural Networks with an R^2 of 0.817. The study demonstrates that ensemble methods like Random Forest can effectively capture complex relationships between professional attributes and salary levels while maintaining good generalization. This research contributes to the field by providing insights into model selection for salary prediction and highlighting the importance of appropriate feature engineering for high-dimensional categorical data.

2 Introduction and Background

Accurate salary prediction is becoming increasingly important for both organizations and individuals in today's dynamic job market. Organizations require these models for competitive compensation structures, and professionals need insights driven by accurate and up to date data about their value in a current market. However, developing accurate salary prediction models presents several challenges due to the complex relationships between various factors that influence compensation. The primary challenge that has kept this solution from being reached already lies in effectively processing and utilizing multiple types of input features - from quantitative measures like years of experience to categorical variables like job titles and education levels. Traditional linear models often struggle to capture these complex relationships, while more sophisticated models risk overfitting when dealing with high-dimensional categorical data that isn't always something that can be generalized to a broader market. The current literature on salary prediction has largely concentrated on industry-specific or regional analyses

with constrained feature sets. Research already done on this through the industry will be discussed in the next sections, particularly by Matbouli and Alghamdi who explored machine learning applications for salary prediction in Saudi Arabia, yet remained geographically limited. We aim to address this clear research gap that exists in developing broadly applicable models that maintain prediction accuracy across varied professional domains while handling diverse job categories effectively that can be applied to a larger scale. Our research will examine multiple modeling methodologies using a comprehensive cross industry dataset that also isn't as limited to just one geographical area either. The research focuses particularly on two critical challenges: managing the dimensionality of categorical variables such as job titles, and accurately modeling non-linear relationships between professional attributes and compensation. Through comparative analysis of modeling approaches and dimension reduction techniques, our research aims to determine optimal prediction strategies that balances complexity and generalization to create a model that can offer key research focused contributions such as the analysis of machine learning methodology effectiveness in salary prediction, the development of techniques for processing high-dimensional categorical compensation data, the examination of professional and demographic factor impacts on salary determination, and the creation of a framework for evaluating salary prediction model performance, while also being of practical use in which we have aimed to provide a comprehensive approach that provides a solution with an end goal of real world applicability. Through careful data preprocessing and model selection, we demonstrate that machine learning techniques can effectively bridge the gap between theoretical research and practical implementation. Our methodology not only advances the academic understanding of salary prediction but also provides organizations and professionals with a more reliable framework for making informed compensation decisions. By addressing both the technical challenges of model development and the practical needs of users, our research provides a study of machine learning methodologies while simultaneously creating a practical salary prediction tool.

3 Literature Review

For our prior research, we looked at the article "Statistical Machine Learning Regression Models for Salary Prediction Featuring Economy Wide Activities and Occupations" by Yasser T. Matbouli and Suliman M. Alghamdi. The article talks about "means of annual salaries across economic activities" [1] in Saudi Arabia using a training dataset with features such as "age group, ethnicity, gender, size of the organization, and type of economic activities" [1]. As for the models, they implemented "artificial neural networks (ANN), tree-based regression (TBR), support vector regression (SVR), and Bayesian-based ML using the Gaussian process regression" (Matbouli and Alghamdi). They also implemented "multiple linear regression (MLR)" [1]. In the end, they concluded that artificial neural networks were the best model for predicting mean annual salary in Saudi Arabia. Some unresolved issues that the current research will address include how to implement one-hot-encoding to process labeled data instead of numeric data during data preprocessing and how to implement Principal Component Analysis (PCA) to reduce high dimensions in the data during the feature selection process. For an

overview of the research, we will first describe the dataset. Then we will do data preprocessing by implementing one-hot encoding. Then for feature selection we will implement principal component analysis. Then we will use that data to train various models such as a linear regression model, a polynomial regression model, a neural network, and a Random Forest model. Then we will conclude which model is the best fit for the data by comparing various evaluation metrics such as mean squared error (MSE), R^2, variance, and bias. Finally, we summarize our findings.

4 Dataset Description and Exploratory Analysis

4.1 Dataset Description

The dataset we used in this study is focused on prediction salaries based on the person's job title and several demographics. The dataset has 6,684 data points and 8 input variables: age, education level, gender, job title, years of experience, country and race. Its output variable is salary.

Age: A numerical feature ranging from 21 - 62 years old with the mean being 33 years old.

Education level: A categorical feature representing the highest level of education attained. The values range from 0 to 3, which represent high school graduate, bachelor's degree, master's degree, and PhDs. Most common degree of education is a bachelor's degree, followed by master's degrees, PhD, then high school graduate. This indicates the education levels of employees in the workforce.

Gender: A binary categorical feature representing the gender of the individual. There are 55% male and 45% female in the dataset.

Job Title: A categorical feature and defines the role of the individual within the organization. There were 129 unique job titles in the dataset. We grouped these job titles into 8 larger categories to decrease the complexity of our model when we do one-hot-encoding.

Years of Experience: A numerical feature representing the number of years the individual has worked in the field. The value ranges from 0 to 34 years with the mean of 8 years.

Salary: The target variable representing the annual salary in US Dollar. The salary ranges from \$350 to \$250,000 with the mean being around \$120,000.

Country: A categorical feature and represents the country where the individual is working.

Race: A categorical feature representing the ethnicity of the individual.

Senior: A numerical feature that has a binary value indicating whether the individual has a senior position.

4.2 Exploratory Analysis

We plotted histograms for the numerical data [Age, Years of Experience, Senior, Education Level and Salary to visualize the distribution of the data as seen in Figure 1. There is a right skew for both Age, Years of Experience, and Senior which indicates there are more young and less experienced employees in the dataset rather than older and more experienced employees. The histogram for Education Level indicates that the majority of the employees in the dataset hold a Bachelor's Degree. There are also a portion that hold a Master's Degree while very few have a PhD or only a High School education. The salary data has a normal distribution. This indicates that most people have around 50,000 to 200,000 dollars in salaries while few people earn

below or above that amount. We used a box plot to handle the outliers in the data. Taking care of these outliers is important because they can prevent the model from learning the underlying patterns in the data. The median age is around 30, with most individuals concentrated at between 20 and 40 years old. The median for Years of Experience is around 10 years with most individuals having between 4 and 20 years of experience. By calculating the quantile of the data, we determine the upper and lower line of the box plot. Any data that lay above the top line and data that lay below the bottom line are considered outliers. The boxplots are shown in Figure 2. Since there were only around 100 outliers found in Age and 70 outliers found in Years of Experience while there are thousands of data points in the dataset, we decide that it is safe to remove these outliers. We replaced the values of the outliers with the values of the lower and upper bound. When we plotted the boxplot for salary in Figure 3, the median salary is around $100,000 with most people having around 50,000 to 200,000 dollars in the library. When we calculated the outliers for salary, there are no data points that are outside of the top and bottom bound of the boxplot.

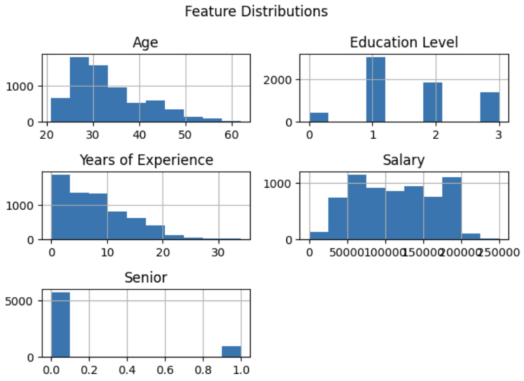

Figure 1: Histograms for numerical features in the model (Age, Education Level, Years of Experience, Salary, and Senior)
Description: There is a right skew for both Age, Years of Experience, and Senior. Education level is highest for 1, which is a Bachelor's Degree. Salary is normally distributed.

5 Proposed Methodology

5.1 Data Preprocessing

During our analysis of the dataset, we observed that many of the features had different scales, which could adversely affect the performance of the machine learning algorithms. For instance, the Education Level feature had values ranging from 0 to 3, while Age spanned a much broader range of 21 to 62. To address this, we applied Z-score normalization, which transforms the data so that each feature has a mean of 0 and a standard deviation of 1. This step ensures that all features contribute equally to the model, preventing features with larger numerical ranges from dominating the learning process. Z-score normalization is particularly important for techniques such as Principal Component Analysis (PCA) and models like Neural Networks (NN), which are sensitive to the scale of the input data. Additionally, we identified and addressed outliers in the dataset. After conducting an outlier analysis, we found approximately 100 outliers in a dataset of 6,684 data points. Since the number of outliers was relatively small compared to the overall dataset, we decided it was appropriate to cap the value of these outliers to the lower or upper bound. This helped improve the stability and accuracy of our models.

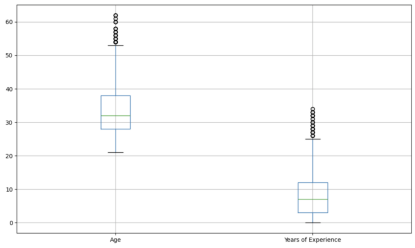

Figure 2: Boxplots for feature Age and Years of Experience.
Description: The data points that are above the top lines of the boxplots are the outliers in the data.

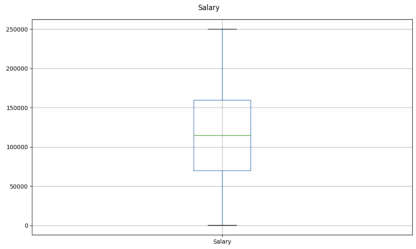

Figure 3: Boxplot for target variable salary.
Description: There are no data points below or above the data, there are no outliers.

5.2 One-Hot Encoding

The dataset contains several categorical variables, such as Job Title, Country, and Race, which are represented as string values. In order to make these variables usable by machine learning algorithms, we applied one-hot encoding. This technique converts each category into a binary vector, where one element is set to 1 to indicate the presence of a category, and the rest are set to 0. However, we encountered a challenge with the Job Title feature, which had 129 unique categories. Applying one-hot encoding to this feature resulted in a high-dimensional, sparse dataset where most of the encoded values were zero. High-dimensional data can increase the model's complexity, potentially leading to overfitting and worse model performance. To mitigate this issue, we grouped the Job Titles into broader categories, reducing the number of unique values from 129 to just 8: Technology, Human Resources, Information Technology, Social Media, Design, Research and Science, and Miscellaneous. This transformation not only reduced the dimensionality of the dataset but also decreased sparsity, making the model more manageable and computationally efficient.

5.3 Feature Selection

In addition, we also employed Principal Component Analysis (PCA) as a dimensionality reduction technique. PCA is effective in transforming high-dimensional data into a lower-dimensional representation, while preserving the most significant features. This helps simplify the model, reduce computational complexity, and potentially improve model performance. Since PCA is sensitive to outliers and missing values, we ensured that these issues were addressed. Before applying PCA to the data, we made sure to scale the data. Then after applying PCA and doing some experimentation, we found that keeping 18 principal components gave us the best results for our models. Additionally, the other principal components had lower variance ratios compared to the rest of the principal components and the total variance ratio did not decrease significantly after dropping them as seen in Figure 4.

5.4 Model Selection

To identify the most effective model for predicting salary in our dataset, we compared four

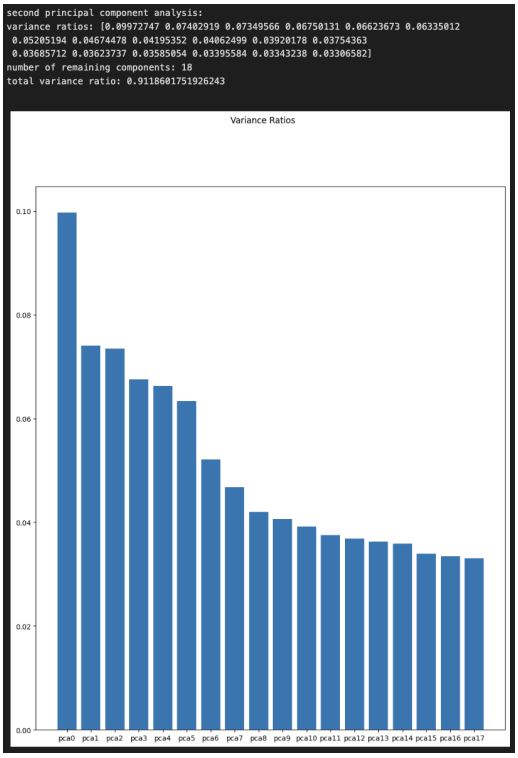

Figure 4: A description and a plot of the explained variances.
Description: The total explained variance ratio after keeping 18 principal components is approximately 91%.

different machine learning algorithms: Linear Regression, Polynomial Regression, Neural Networks, and Random Forest. Linear Regression assumes a linear relationship between the input features and the target variable (salary). While this model is simple and interpretable, it is unable to capture more complex, non-linear relationships in the data. As a result, Linear Regression may not perform well when the true underlying relationships between features and salary are more intricate. Polynomial Regression models the relationship between the input features and the target as an n-th degree polynomial. Transforming input features into polynomial features allows the model to capture non-linear relationships. We experimented with polynomial degrees of 1, 2, and 3 to assess how higher-degree polynomials affected model performance. While higher-degree polynomials allow for more flexibility in capturing non-linear patterns, there is the risk of overfitting if the model becomes too complex for the data. Neural Networks are capable of modeling complex, non-linear relationships by using activation functions that introduce non-linearity into the model. However, we faced a challenge due to the magnitude of salary values, which caused the model to have difficulty converging during training. To address this, we applied scaling to the target variable (salary) before training the network, and then reversed the transformation after making predictions. This scaling process helped the network learn more efficiently and produced more stable predictions. Random Forest is an ensemble learning method that builds a collection of decision trees using bootstrap sampling (random subsets of the data). Each tree is trained on a different subset of the data, and a random subset of features is considered when splitting each decision node. After training, the predictions of all trees are aggregated to make the final prediction. This approach helps prevent overfitting and increases the model's generalizability. Random Forest is particularly effective at modeling complex, non-linear relationships and performs well with minimal tuning, making it a strong candidate for this problem.

5.5 Data Splitting and Training

We split the data into training and test sets, using a 75% training and 25% test split. The training set was used to train the models, while the test set was reserved for evaluating the model's performance. This division ensures that the models are tested on unseen data, providing a realistic estimate of how they would perform in a real-world scenario. After training the models, we assessed their performance by comparing the predicted salaries to the actual salaries in the test set. We used several metrics to evaluate the models' performance. Mean Squared Error (MSE), which measures the average squared difference between the predicted and actual values. A lower MSE indicates that the model's predictions are closer to the true values, signifying better performance. R^2, which indicates the proportion of the variance in the target variable (salary) that is explained by the model. A higher R^2 score means that the model explains a larger portion of the variance in salary, reflecting a better fit to the data. Additionally, we looked at bias and variance. Bias measures the error introduced by the model's assumptions or oversimplifications. High bias typically leads to underfitting, where the model is too simple to capture the underlying patterns in the data. On the other hand, variance measures how much the model's predictions would change if it were trained on different subsets of the data. High variance can lead to overfitting, where the model becomes too complex and starts to capture noise or random fluctuations in the training data. The goal is to find a balance between bias and variance, ensuring that the model is neither too simplistic (underfitting) nor too complex (overfitting). By using these metrics, we assess each model's predictive accuracy and its ability to generalize well to new data.

6 Experimental Results and Evaluation

We implemented a total of four models which consisted of a Linear Regression Model, a Polynomial Regression Model, a Neural Network, and a Random Forest Model. To compare the predictions of our models, we used evaluation metrics such as bias, variance, mean squared error, and R^2.

6.1 Linear Regression

For ‘ Linear Regression model, the mean squared error was 843640750.4135907. The mean squared error was very high relative to the rest of our models and may indicate that the Linear Regression model underfitted the data. The R^2 value was 0.6957691054230675. This can indicate a somewhat strong relationship between the input features of our dataset and salary. The variance was 1947812325.7402577 which was lower than the variance for our polynomial regression model. The $bias^2$ was 2773515869.94308 which was lower than the $bias^2$ for our polynomial regression model.

6.2 Polynomial Regression

For our Polynomial Regression model, we tested degrees 1, 2, and 3. The results of the evaluation metrics can be found in Figure 5. For higher degrees such as 2 and 3, the $bias^2$ and variance increased significantly as seen in Figure 6. Since the total error of the model is the sum of $bias^2$, variance, and the irreducible error of the model, this means that the total error of the model increases significantly for higher degrees. Addi-

tionally, for higher degrees, the mean squared error also increases significantly as seen in Figure 7 and the R^2 is negative as seen in Figure 8. This implies that the relationship between the input features of our dataset and salary is not strong when modeled by our Polynomial Regression model compared to our other models. It means that a horizontal line would be a better fit for the data than our Polynomial Regression model at higher degrees.

Degree	MSE	R^2	Variance	Bias^2
1	843640750.4135906	0.6957691054230676	1947812325.7403	2773515869.9431
2	710575134598.6365	-255.2452190189802	714527327310.6324	3030568001.9752
3	2.3777742691232093e+17	-85746496.26708035	237738946227114784.0000	38052142807570.6250

Figure 5: Evaluation metrics for the Polynomial Regression model.
Description: For higher degrees, the Polynomial Regression model did not perform well.

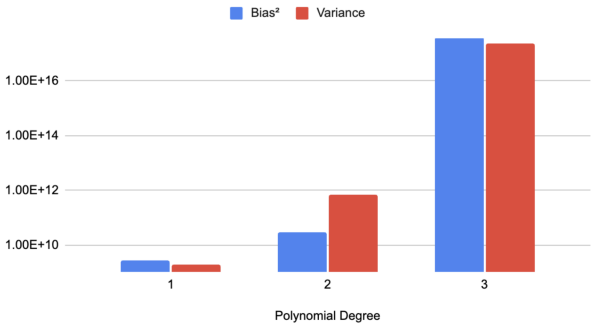

Figure 6: A graph of the $bias^2$ and variance for the Polynomial Regression model.
Description: For higher degrees, the $bias^2$ and variance increased significantly.

6.3 Neural Network

For our Neural Network, the mean squared error was 501284753.93979967 which is lower than

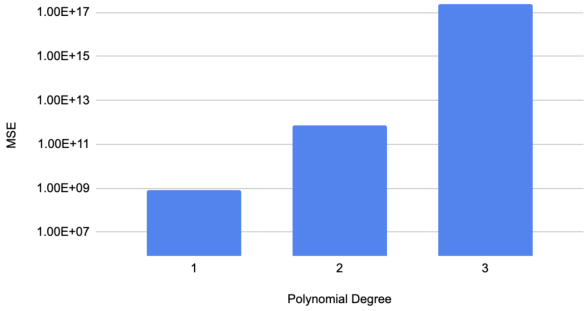

Figure 7: A graph of the mean squared error for the Polynomial Regression model.
Description: For higher degrees, the mean squared error increased significantly.

our Linear Regression model and the Polynomial Regression model. The R^2 value was 0.8192283752840089 which indicates a stronger relationship between the input features of our dataset and salary.

6.4 Random Forest

For our Random Forest model, the mean squared error was 422546711.4490019 which was lower than the rest our models. The R^2 value was 0.847622623774813 which was higher than the rest of our models and indicates the strongest relationship between the input features of our dataset and salary. These evaluation metrics are the best out of all the models which indicates that the Random Forest model is the best fit for our dataset. This makes sense because our dataset had a high number of features and a Random Forest model is a good fit for datasets with high number of features. This is because the Random Forest model selects a random subset of features instead of selecting all of the features during the development of each decision tree.

To conclude our experimental results, figures

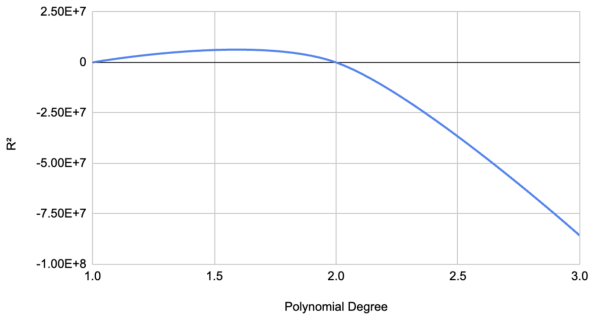

Figure 8: A graph of the R^2 value for the Polynomial Regression model.
Description: For higher degrees, the R^2 value becomes negative.

9, 10, and 11 compare the evaluation metrics of each model.

Model	MSE	R^2
Linear Regression	843640750.4135907	0.6957691054230675
Polynomial Regression (Degree 1)	843640750.4135906	0.6957691054230676
Polynomial Regression (Degree 2)	710575134598.6365	-255.2452190189802
Polynomial Regression (Degree 3)	2.3777742691232093e+17	-85746496.26708035
Neural Network	501284753.93979967	0.8192283752840089
Random Forest	422546711.4490019	0.847622623774813

Figure 9: A table of the mean squared error and R^2 value of each model.
Description: Our Random Forest model had the lowest mean squared error and highest R^2 value.

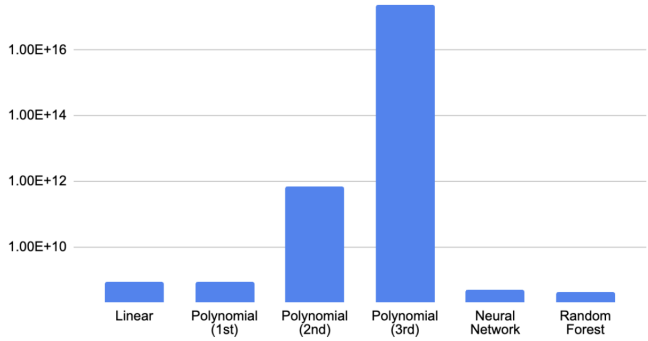

Figure 10: A graph of the mean squared error for each model.
Description: Our Random Forest model had the lowest mean squared error.

7 Conclusion and Discussion

The results of the study tell us that characteristics of the dataset play a key role in determining what data preprocessing needs to be done and what feature selection needs to be done. Our study used a dataset with high dimensions. Due to this, we needed to broaden the job titles during the preprocessing phase in order to reduce error in our models and to get the models to converge. We also needed to implement one-hot encoding since the dataset contained not only numerical data but also categorical data. In addition to this, we needed to use Principal Component Analysis to reduce the high dimensions during feature selection. The characteristics of the dataset also determine what model is the best fit for the dataset. The Random Forest model was the best fit for our dataset because it takes into account the high dimensions of our dataset when trying to find the feature with the highest information gain to split a decision node on.

8 Github Link and Project Roadmap

8.1 Github Link

https://github.com/ruohan8/ECS171Project

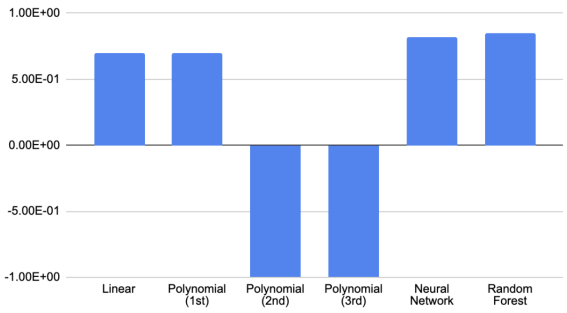

Figure 11: A graph of the R^2 value for each model.
Description: Our Random Forest model had the highest R^2 value.

8.2 Contributions

Ruohan Huang: Exploratory Data Analysis, One-hot encoding, Linear Regression model, Polynomial Regression model, Random Forest Model, flask, Methodology section
Jacob Nguyen: broadened job titles, correlation matrix heatmap, principal component analysis, literature review, conclusion, wrote final report in LaTeX
John Sebastian Solon: Dataset description, feature selection, broadened Job Titles, code debugging and fixing, rewrote structure of ipynb to be more readable and easier to debug, testing, mid-quarter report and final report
Frank Wem Guang Zhu: Implemented the Neural Network model, did data research and created a skeleton of the front-end, Prepared the project documentations (final and check in reports).
Amritpal Zenda: Project proposal, dataset research and selection, attribute selection, mid-quarter report, model testing, code debugging, final report

8.3 Roadmap

Week 1-2: Data Preprocessing and Exploratory Data Analysis (EDA)
Week 3: Feature Selection and One-Hot Encoding
Week 3: Mid-Quarter Project Progress Report
Week 4-7: Model Construction
Week 8-9: PCA, Model Evaluation
Week 10: Final Project Report, Demo, and Source Code

References

[1] Matbouli, Y. T., & Alghamdi, S. M. (2022). Statistical Machine Learning Regression Models for Salary Prediction Featuring Economy Wide Activities and Occupations. Information, 13(10), 495. https://doi.org/10.3390/info13100495