Brady Berg

Fall 2021

**Prediction and Inference of STEM Salaries**

**Abstract and Introduction**

With the extremely high level of competition between universities and institutions, sought out students and individuals find themselves in a high stress, highly intellectual industry where competition is fierce. Many students count degrees of graduate schooling and internship years on their resumes, which are often their ticket to certain competition themselves. As students--of any age--are encouraged to attend higher level university education and pursue increasingly furthering careers in their field of choice, a strong understanding of what is considered 'fair pay' within your industry is important. With people from many well-known, highly reputed institutions, there is still the reality that the number of years one spends in any position and the level of education they have obtained may not be the only factors determining compensation. This report takes a closer look into the role these attributes play in compensation within the STEM environment. An analysis of 4,608 models of varying methodology shows that these data and models can explain about 77% of the variance in salary as well as the factors that contribute towards it.

**Related Work**

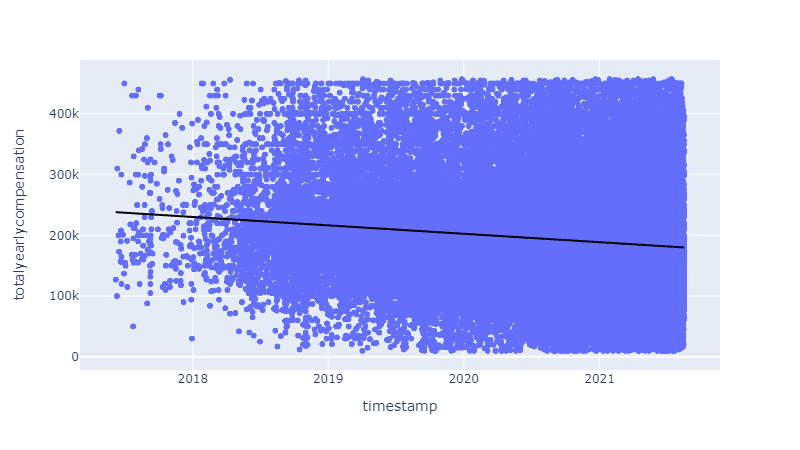
The majority of related efforts on this dataset pertain to exploratory data analysis and visualization. While these are important to understand the structure and high abstraction nature of the data, they are restricted to comparisons between two or three variables at a time and can be interpreted differently between viewers. Other methods include a statistical t-test to evaluate the significance of a PhD on salary as well linear correlation and an attempt at k-means clustering. In the next section an overview of the dataset is provided before presenting the methodology and results of this project which provides more concrete answers to these questions.

**Data**

|  |  |  |
| --- | --- | --- |
| **Feature** | **Number of Unique Values** | **Number of Unique Values After Binning** |
| **Timestamp\*** | 59045 | 0 |
| **company** | 1623 | 1076 |
| **level** | 2832 | 591 |
| **title** | 15 | 15 |
| **totalyearlycompensation** | 463 | 463 |
| **location** | 1041 | 379 |
| **yearsofexperience** | 49 | 49 |
| **yearsatcompany** | 66 | 66 |
| **tag** | 2937 | 307 |
| **Basesalary\*\*** | 402 | 0 |
| **Stockgrantvalue\*\*** | 382 | 0 |
| **Bonus\*\*** | 274 | 0 |
| **gender** | 4 | 4 |
| **Race** | 5 | 5 |
| **Education** | 5 | 5 |
| **State\*\*\*** | 50 | 50 |

***Table 1: Features of the dataset and their number of unique values after removing outliers and extraneous data. Totalyearlycompensation is response. \* = Removed, \*\* = Removed due to direct correlation with response, \*\*\* = Computed feature.***

The dataset after removing irrelevant and disaggregated response features is mainly categorical. Therefore, the number of classes in each feature are evaluated to see possible problems due to many observations of underrepresented classes. As a solution, categories which have less than 5 observations are labeled as “Other” under their respective feature. This helps prevent overfitting in the modeling process and would aid in limiting the number of features in a one-hot encoded transformation. Here we also justify removing the timestamp variable.



*Figure 1: Annual Compensation vs. Timestamp. R^2 = 0.016.*

There seems to be no change in salary over time in this data. This leads to an unintended benefit. There are two features left in the dataset which are numerical and ordinal: years of experience and years at company. Although they are ordered features, they are not continuous with ½ year granularity. In addition, they are highly correlated, so one of them may be removed in the feature selection process anyway. By treating these variables as categorical, we can transform and model them in the same space which will be described later.

A picture containing text, vector graphics

Description automatically generatedAnother interesting aspect in this data is salary by state.

*Figure 2: Annual Salary by state*

As is it evident, California has a much average salary than the other states. The Midwest and southern states have a lower salary.

**Methods**

As stated earlier, the remaining variables in this dataset are treated as categorical. In this sense, compatibility issues are avoided with respect to some methods which require the same data type (categorical or continuous) such as PCA. On the same note, the first step of every one of the 4,608 models was applying a target encoder. After this step, all features are in the same space, and a large number of features is avoided from one hot encoding. The following table explains the methods applied to each step of the modeling pipeline in this work.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Scaling** | **Feature Engineering** | **Feature Selection** | **Dimensionality Reduction (Number of PC)** | **Model** |
| None | None | None | None | Linear Regression |
| Standard | Polynomial Features (degree = 3) | Variance Threshold (0.5) | PCA (10) | Stochastic Gradient Decent Regressor |
| Robust | Spline Transformer | Select K Best (k=11) | PCA (6) | Random Forest Regressor |
| Quantile Transformer | Feature Agglomeration | Select K Best (k=5) | PCA (3) | Gradient Boosting Regressor |
| Min Max |  | Select Percentile (0.9) |  | Histogram-based Gradient Boosting Regression Tree |
| Power Transformer |  | Select Percentile (0.7) |  | Multi-Layer Perceptron Regressor |
|  |  |  |  | Ada Boost Regressor |

*Table 2: Transformation and modeling methods used.*

To find a sutiable model, every permutaion across columns of the above table were fit and cross validated. Importantly, each step in the pipeline has a passthrough option to allow the data to pass through the step without transforming so as to avoid transformations when they are not beneficial. Table 2 shows when hyperparameters were specified. Otherwise, the default values in the sklearn python library were used. Hyper parameter tuning for a specific model can be a topic of further research given the runtime of about two hours everytime a set of steps of this size are evaluated with the resources available in this project.

Every combination of the columns above are inserted as steps in a modeling pipeline, fitted, and evaluated across 2 validation folds on their squared error. Lastly, the resulting models are examined.

**Results**

The best estimator from this study is a model which follows the following pipeline: Target encoding, Min Max Scaler [0, 1], Spline Transformer with 5 knots, then fit to a Histogram-based Gradient Boosting Regression Tree with a learning rate of 0.1.

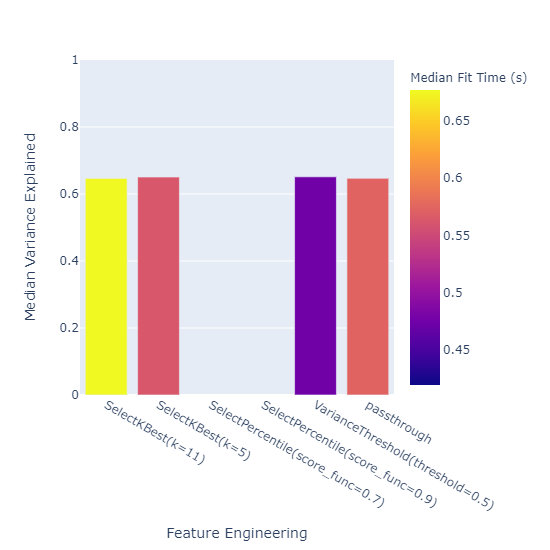
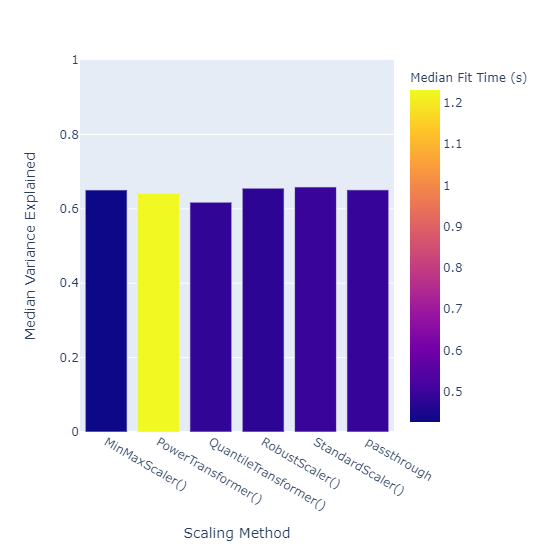
Graphical user interface, text

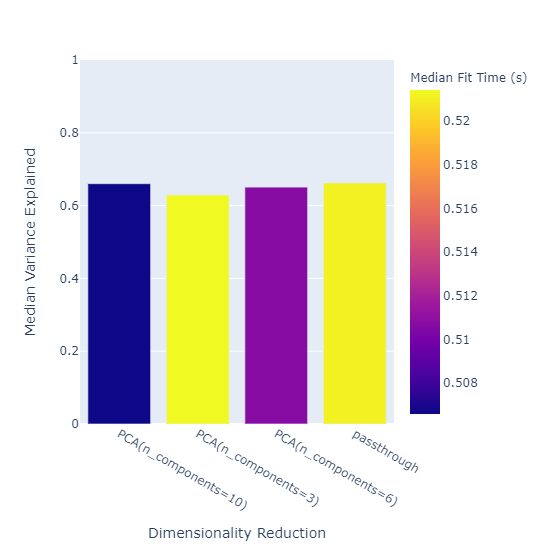
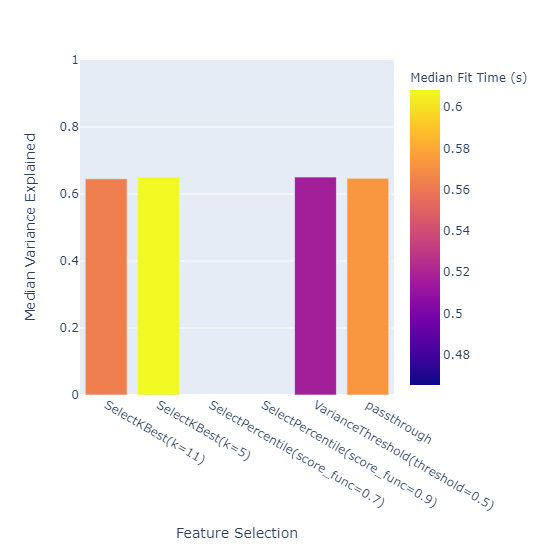
Description automatically generated

*Figure 3: Resulting best model in terms of explained variance.*

This model is the result of permutations where both feature selection and dimensionality reduction were not used. This is sensible given that tree methods have an inherent feature selection property.

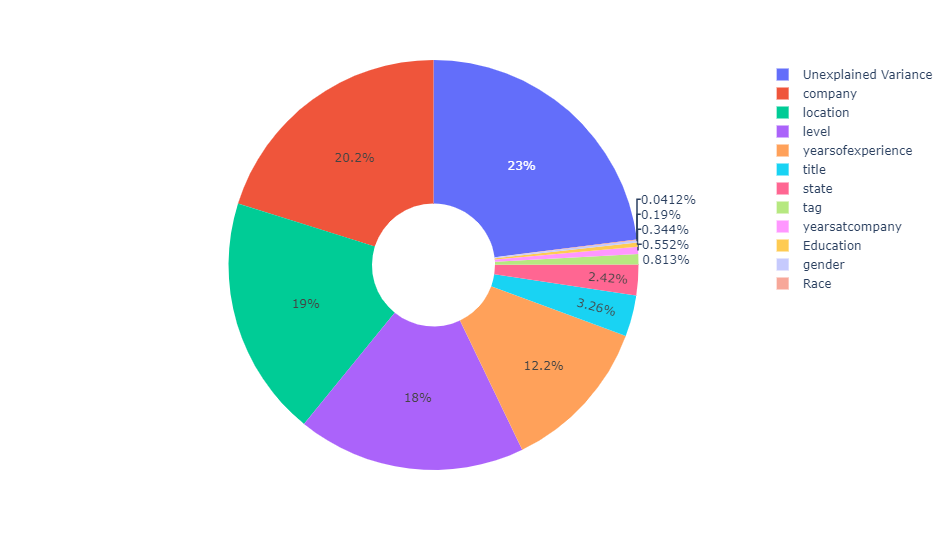
Next, further insight is derived from the median R^2 value over models of each possible method in each step of the pipeline. Median fit time is also included for efficiency considerations.





*Figure 4: Median variation explained by each method in the pipeline.*

Noticeably, there is not a significant difference between the methods. However, there are some methods which failed to fit completely. Lastly, feature permutation feature importance is estimated across 10 permutations for each variable. It is important to note that these are an estimation given that an assumption of non-multicollinearity cannot be met.



*Figure 5: Feature Importance and unexplained variance.*

Figure 5 suggests that years of experience and education, which many believe to be important factors in salary, explain only a small fraction of the variance (0.122 and 0.003 respectively). Assuming this data and model are accurate, the most significant factor is the specific employer and location. Including unexplained variance in the figure also highlights that there are uncaptured variables that contribute to salary not present in this data or that these variables are not captured completely in the model.

**Conclusion**

This project used a brute-force methodology to choose a model salary as a function of the other features in the dataset. While this may have been a viable option, further hyperparameter tuning could be used to optimize the model. In addition, correlation between variables was not considered, and a feature (state) was added which was directly derived from location. While randomized tree methods are known to be robust against Multicollinearity, permutation feature importances are not. Further work could consider this in addition to implementing conditional permutation to avoid extrapolation.

The results shown suggest that the public perception of what contributes to how much someone is paid may be skewed. Here, company and location show the most importance in our model which captures 77% or the variance in salary. More work on this subject could produce more robust methods with more credible data.

Works Cited

deepguydeepguy 1, et al. “Multicollinearity in Decision Tree.” *Data Science Stack Exchange*, 1 June 1966, https://datascience.stackexchange.com/questions/31402/multicollinearity-in-decision-tree.

Ogozaly, Jack. “Data Science and STEM Salaries.” *Kaggle*, 10 Oct. 2021, https://www.kaggle.com/jackogozaly/data-science-and-stem-salaries.

Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., … others. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.