Salary Prediction for Data Science and Related Jobs

Alex Ojemann	Tyler Kimbell
Department of Computer Science	Department of Computer Science
University of Colorado Boulder	University of Colorado Boulder
Boulder CO, USA	Boulder CO, USA
Alexander.Ojemann@colorado.edu	Tyler.Kimbell@colorado.edu

Introduction

When thinking about applying for a job or starting a career, it is important to understand how different positions, skill sets, and geographical location will affect someone's salary. With many different job postings it would be beneficial to give an estimate of an average salary given some rudimentary data. Stakeholders include students looking at majors they want to go into, people planning out careers, and people looking for a job.

Literature Survey

Our data set contains 742 entries of data science and related jobs from glassdoor. One publication that explores the same data set explores a number of techniques including Ridge and Lasso regression, random forest, support vector regression, gradient boosting, and adaboosting, along with a dual adaboosting system of their design[1]. The authors use min-max scaling to preprocess their numeric variables and transform all categorical variables into binary dummy variables for each category using one hot encoding. They use adaboosting feature importance in order to limit the excessive dummy variables, filtering out any with a feature importance of less than 0.001. They ultimately evaluate the root mean squared error on their training set, the root mean squared error on their test set, and the mean absolute error on their test set for each of their models. The adaboost and dual adaboost models perform the best but exhibit a high degree of overfitting given the large split between their training root mean squared error and testing root mean squared error. Their accuracies can be used as a reference point for the accuracies of our model because they use the same data set.

[1] “Salary Prediction Based on the Dual-Adaboosting System | Semantic Scholar.”

One of the challenges of our data set is that the categorical features have many distinct values. For example, one of the features in our data set is the job title, which has 264 unique values. If we were to one hot encode this variable, we would get 263 binary dummy variables. This subjects us to the curse of dimensionality, which is the effect that large numbers of features leave you more vulnerable to finding trends in your data that aren't generalizable[2]. Two frequently used tactics for addressing this are Ridge[3] and Lasso[4] regression, which penalize feature weights in a regression model to limit the number of features for which the optimal model will have a nonzero coefficient. While these methods maintain the interpretability of the model, they may limit accuracy. Principal component analysis, or PCA, compares favorably to other methods for encoding variables in terms of maintaining accuracy in some applications with high dimensional categorical data[5]. However, this method sacrifices interpretability as PCA creates features using orthogonal linear components of all the variables that explain the most variation possible.

One of the other challenges that occurs when creating a model to predict salaries is hyperparameter optimization. Many machine learning models require hyperparameters that are computationally expensive to estimate with grid search and can lead to a suboptimal model if manual trial and error doesn't arrive at an optimal value. Randomized parameter optimization avoids the computational expense of grid search but is less prone to missing optimal combinations than manual exploration[6].

Proposed Work

As previously mentioned, the dataset we will use for our project is located on Kaggle. To preprocess the data, we first analyze each of the features and determine if there are significant outliers that need to be removed. Next we will fill NA data with the mean for numeric features or the mode for categorical features. Then we will perform min max scaling on all of the numeric features. For the categorical features we plan to take a few different approaches to encoding them and comparing the results. One is to one hot encode them which will result in thousands of dummy variables and put them into ridge and lasso regression models to regularize the features. This approach is similar to that of the first paper mentioned in the literature survey but we will let the ridge and lasso regression models choose from all of the features rather than using adaboost beforehand. This approach could still be subject to the curse of dimensionality in that one of the many dummy variables could appear to be significant in the training data so the model includes it but it may not generalize. For example, the top feature in ADABoost feature importance was "GRM Actuarial", a dummy variable only found in three entries. Another

[2] Debie and Shafi, "Implications of the Curse of Dimensionality for Supervised Learning Classifier Systems."
[3] Marquardt and Snee, "Ridge Regression in Practice."
[4] Tibshirani, "Regression Shrinkage and Selection via the Lasso."
[5] Farkhari et al., "New PCA-Based Category Encoder for Cybersecurity and Processing Data in IoT Devices."
[6] Bergstra and Bengio, "Random Search for Hyper-Parameter Optimization."

approach will be to generate principal components for the dummy variables and use these as features in our models along with the numeric variables. The number of principal components will be decided later. The final approach will be to attempt to group these categorical variables into a manageable number of dummy variables representing characteristics of the variable. For example, the headquarters variable which contains the location of the company headquarters could be grouped into urban or rural and a dummy variable could be made for whether the description contains a defining word about an industry. We plan to put these PCA features and manually engineered features into a random forest, an xgBoost model, and an adaboost model using randomized search to optimize hyperparameters, then compare the performance of each of these models along with the ridge and lasso regression models with all of the one hot encoded variables.

Evaluation

To evaluate each of our models, we plan to use root mean squared error and mean absolute error, each of which are defined below.

Root Mean Squared Error = (sum(y - y_pred)/n)^½
Mean Absolute Error = (sum(abs(y - y_pred))/n)

We chose these evaluation metrics because they allow for direct comparison to the models in the first citation of the related work section. We plan to evaluate each of these metrics on each model using 10-fold cross validation to avoid the arbitrary nature of using one single train test split.

Milestones

- Our first milestone will be to perform the outlier analysis, NA filling, and numeric scaling described in the proposed work section. This will be done by October 15th.

- Our second milestone will be to complete the encodings of categorical features described in the proposed work section. This will be done by November 9th, along with the project progress report.

- Our third milestone will be to train our models, optimize hyperparameters, and compare using our evaluation metrics. This will be done by December 7th, along with the final report.

References

Bergstra, James, and Yoshua Bengio. “Random Search for Hyper-Parameter Optimization,” n.d.
Debie, Essam, and Kamran Shafi. “Implications of the Curse of Dimensionality for Supervised

Learning Classifier Systems: Theoretical and Empirical Analyses.” *Pattern Analysis and Applications* 22 (May 1, 2019). https://doi.org/10.1007/s10044-017-0649-0.

Farkhari, Hamed, Joseanne Viana, Luis Miguel Campos, Pedro Sebastiao, and Luis Bernardo. “New PCA-Based Category Encoder for Cybersecurity and Processing Data in IoT Devices.” arXiv, May 23, 2022. http://arxiv.org/abs/2111.14839.

Marquardt, Donald W., and Ronald D. Snee. “Ridge Regression in Practice.” *The American Statistician* 29, no. 1 (1975): 3–20. https://doi.org/10.2307/2683673.

“Salary Prediction Based on the Dual-Adaboosting System | Semantic Scholar.” Accessed September 26, 2023. https://www.semanticscholar.org/paper/Salary-Prediction-Based-on-the-Dual-Adaboosting-Chen-Zhan/641c323bce27ea7f1791f176991a3c10df5d0f70.

Tibshirani, Robert. “Regression Shrinkage and Selection via the Lasso.” *Journal of the Royal Statistical Society. Series B (Methodological)* 58, no. 1 (1996): 267–88.