Are teachers being compensated fairly?: A comprehensive analysis and predictive modeling of teacher salaries in Illinois

Fiona Baenziger, Brittany Ciura, Kyle Spanski, and Alison Korhan
June 10, 2019

Executive Summary:

Fairness is when something is considered to be impartial and free from self-interest, prejudice, or favoritism. Fairness applied to salary is a different, more complicated application of the term. People can have opposing opinions when discussing the idea of fairness when it comes to money but in an idealistic world, compensation is influenced by actual merit and experience and not with an inherent bias. Each professional field has their biases that influence how salary is determined and this concept of fairness is different from field to field. Our team has been tasked with creating a model to determine if there is any present bias in teacher salaries and if they are being compensated fairly by analyzing the predictive characteristics. This includes a full comprehensive analysis of teacher salaries in Illinois, as well as creating a predictive model that will help generate future salaries.

The State of Illinois provides public salary records of primary and secondary teachers that span an entire decade, including salary, school, experience, and demographic information. From the data set, we can analyze if there are certain features or demographic characteristics which predict higher salaries. In addition, the State of Illinois also provides public test score records from schools across the state that include information pertaining to Reading and Math Standardized test scores and the ratio of teachers to students. With this, we can compare additional factors that may contribute to teacher compensation and start to question whether this is a good metric for salary.

To assess which characteristics contribute the most to teacher salary, our team conducted a variety of feature selection techniques to determine those characteristics with the highest correlations. With these features determined, our team tested the predictiveness of popular machine learning techniques against the dataset. We decided to focus specifically on CatBoost, a model that accounts for the categorical nature of the dataset. After conducting our analysis, the results provided interesting insight to the wage debate. While the original data set showed evidence of salary bias, especially among gender and ethnicity, the model shows a different picture. This demographic information, such as gender and race/ethnicity, are not identified as the ‘most important’ predictors for teacher salary in our final model. This means that while a wage gap still exists based on choice demographics, these are not features that most prominently determine teacher salary. The main factors that contribute to the model are related to education, experience level, and location. These are the features that show a relatively high level of wage inequality. Of the eight selected predictors in the final CatBoost model, five of them are related to education and experience.

The importance of these features in teacher salary leads the assumption that there is a much larger systematic problem that is causing this wage inequality. Factors like school funding and additional benefits, such as pensions, could have a bigger impact on the inequalities in teacher compensation that should be investigated more thoroughly. In order to comprehensively determine if teacher salary is ‘unfair’, the topic requires a more granular level analysis that omits big factors such as education and location. By evaluating fairness in teacher compensation, we can hopefully work towards solving the inequalities across the state of Illinois.

Abstract:

The purpose of this research was to run a comprehensive analysis of teacher salary distribution in Illinois and create an accurate predictive model to determine salaries. Additionally, we also wanted to use the provided data to determine if teachers were being compensated fairly or if there was an inherent bias to any particular groups. The data used was from 2003 to 2012 and from the Illinois Standard Teacher Service Record (TSR). Several feature selection techniques were applied on the data set and eight total features ended up being selected for the final model. After comparing results from Decision Trees, Random Forests, Gradient Boosting, and CatBoost, the CatBoost model was selected as the final model and generated an R^2 of 0.744 on the training data set and 0.734 on the test data set. A CatBoost model was also applied to a separate data set that had test score information by district and this model had an R^2 of 0.8 for the training data set and 0.73 for the test data set. Our hypothesis that teacher salaries were unfair and biased was ultimately disproved through this analysis as the eight main features selected for the model were based on experience and personal merit and did not show an inherent bias.

Introduction:

By definition, the idea of fairness is when something is considered to be impartial and free from self-interest, prejudice, or favoritism. In terms of salary, the concept of fairness is important to determine if compensation is influenced by actual merit and experience or with an inherent bias. Given the growing wage gap between teacher compensation and other professions (Allegretto and Mishel, 2018), determining if there is an impartiality in compensation is even more necessary. Our team has been tasked with coming up with a comprehensive analysis of teacher salaries in Illinois, as well as creating a predictive model that will help generate future salaries. In addition, we were interested in learning whether there was any present bias in teacher salaries and if they were being compensated fairly given the provided data.

For a teacher compensation program to be truly fair, an individual teacher's compensation should not be influenced by demographics such as gender, ethnicity, or other socioeconomic factors. Our main goal in this research is to determine if teachers in Illinois are being compensated purely based on their experience and/or qualifications. For example, a fair compensation structure would mean two teachers with the same experience level and located in the same district would be compensated about the same, regardless of their demographic affiliation.

In order to analyze teacher salaries and identify any biases, our research goal was to create a predictive model to help generate teacher salaries given certain variables. The dataset we were given includes a decades worth of teacher compensation data in Illinois public schools. As the data we were provided only gives information on teacher salary and not any other form of compensation (i.e. vacation days, bonuses, reimbursements for continuing education, etc.), we will not be able to analyze the impact these extra perks have on overall compensation. Generating a 'best' predictive model will allow us to narrow in on which attributes are most impactful in teacher salary and if those attributes allude to an inherent bias or unfairness in what is being used to set compensation levels.

Literature Review:

While research on determining a bias in salaries has long been discussed, we did not find much evidence of previous studies being done to create a predictive model for salaries, especially in the education

industry. Even so, there were a few papers that helped influence our initial process to this problem. One of the closest examples was a study that aimed to predict a student's future salary after they graduated from college (Kongchai et al., 2016). The main goal of this paper was to determine if salary drove a student's motivation to do well in classes. The researchers used a random forest technique to predict salary and were able to achieve a high accuracy, which is a model we also applied to our data set.

Another area we wanted to focus on in the research is what drives salary increases among teachers. One article discovered that a 1% increase in the base salary of a teacher drives up the proportion of teachers hired (Hendricks, 2015). In that same research study, it was discovered that paying one teacher more drives up salaries among all teachers at that school. We plan to investigate this further by utilizing student performance against teacher compensation to support this relationship.

Finally, we referenced a paper (Huang et al., 2019) that focused on how the newer CatBoost model could be used in predictive capabilities that had previously been applied with Random Forests and Support Vector Machines. A CatBoost model was something we wanted to look into for this research as our data set included many categorical variables. These categorical variables can be handled with CatBoost without first transforming them to numerical variables. The goal of the research in the paper was to create the best model to predict evapotranspiration in humid areas in China. In doing so, the researchers compared the performance outputs of a CatBoost model to Random Forests and Support Vector Machines, both of which the CatBoost model was able to outperform.

Data Preprocessing:

Data Source:

The main data set used for this analysis is from the Standard Teacher Service Record (TSR) and includes data from 2003 to 2012 totalling 1,624,887 instances. The information in these data sets is all public record and available online. There are 60 total variables included in the data set including salary amounts for each teacher, which we will be using as our dependent variable. Other notable variables included in the dataset are school information, district information, education levels, both state and district experience, assignments taught, months employed, and position type.

Data Transformation:

Each year of data was provided to us as a separate file so we first merged the data sets together before beginning the preprocessing. From there, we evaluated each of the provided variables to determine which would be useful in our analysis. There were many variables that could be safely removed and these included several that listed out teacher assignments after their first main assignment, which ended up being blank for most teachers anyways. We also removed all specific identifying variables such as teacher names, school addresses, district addresses, and variables for beyond the first three digits of a zip code. Finally, all cases where salary was equal to $0 were removed as well as any cases that had missing values. We started with a large number of instances so removing these cases did not impact our ability to work with the data, removing less than 8% of the data with any missing value for any variable (123204 of 1624887 records were remaining). We also ensured that there was no pattern to the missing values.

Several of the variables in the data set were transformed in order to better work for our models. Location was one variable we predicted would have a large impact on salary, but we were finding many teachers in the Chicagoland area. To help create more separation among location, we created a new variable with three total categories: Chicago, suburbs of Chicago, and the rest of Illinois. These were assigned based on

the first three digits of the district's zip code. This allowed us to still be able to evaluate location without having too much of an imbalance towards teachers in the city. We ended up with around 50% of our data from suburban Chicago districts, 16.6% of our data from the Chicago district and 33.3% of our data from other districts.

For those variables that were categorical in nature, we transformed them into dummy variables so they could be applied to numerical modeling techniques. Some of these variables include ethnicity, gender, zip code type, teacher type (i.e. primary school teachers, high school teachers, etc.), teacher education level, and whether a teacher taught at multiple locations. Additionally, we also created a similar data set where these categorical variables were not turned into dummies as we could apply this to a CatBoost model.

Finally, the most important variable we transformed was for salary, which was to be our main predictor variable. However, after noticing a large proportion of teachers only worked for about ten months out of the year, we transformed this to represent monthly salary instead. Each teacher's yearly salary was divided by the number of months they worked in the year. Once calculated, we also adjusted each monthly salary value for inflation dependent on the year that salary was recorded, based on US CPI measure from the Federal Reserve for each year. This new adjusted monthly salary variable will serve as our predictor variable for the remainder of the research.

Outlier Analysis:

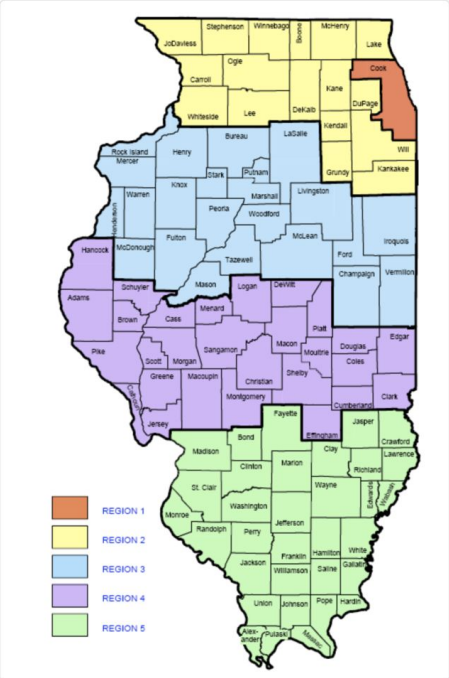
Some of the exploratory analysis that will be discussed in the next section showed there were a few severe outliers in the data. As our dataset included a large number of cases, we first made the decision to remove all those that had a z-score beyond three standard distributions from the mean. However, a more thorough analysis of this 'outlier' group showed us that there might be more information to gain by keeping them in the data set, especially as our final goal is to tackle the issue of fairness. Was this large group being compensated more due to their own merits or a demographic bias? Ultimately, there were eight total outliers that were removed for having a salary very much outside of the dataset and 200,106 salaries that were equal to $0. This left us with a final count of 1,301,569 instances in the data.

Test Scores Data:

There has been a lot of debate over the years about whether students' standardized test performances should influence the salary of a teacher to provide an incentive for better teaching. To explore this idea, we are interested in assessing whether test scores play a role in the determination of salary. The Illinois State Board of Education publicly posts test scores for every school and school district each year, such as grade level reading and math test scores, ISAT scores, and ACT scores. To make the scores more manageable, we calculated the average reading and average math scores across each district. The test scores are aggregated by district so instead of joining the test scores data to the individual teacher salaries dataset where every teacher in a school district would have the same test score, a supplemental Test Scores dataset was created to reduce redundancy. From the teacher salary dataset, we calculated the Median Adjusted Monthly Salary and number of teachers for each school district per year and joined this information with the Test Scores dataset by their common unique identifier RCDT, the Region-County-District-Type Code. The Test Scores dataset also included details on the school district location, number, county and total enrollment. We also wanted to explore if the student to teacher ratio had an impact on teacher salaries, so we created a new feature Students/Teacher by dividing the total enrollment by the number of teachers employed in each school district.

We thought school/district location would be an important feature in our analysis, but some of the models used to analyze the dataset require all of the features to be numeric. A common method to transform

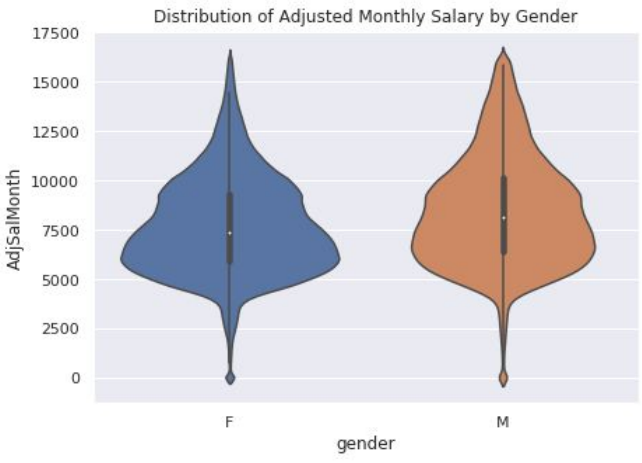
categorical variables is to create dummy variables using one-hot encoding for each value of the category. However, with 102 counties in Illinois, this approach was not suitable for our analysis. We decided to group the counties into five regions to reduce the number of categories. The regions were grouped based on this map, with Region 1 including only Cook County (below in orange).

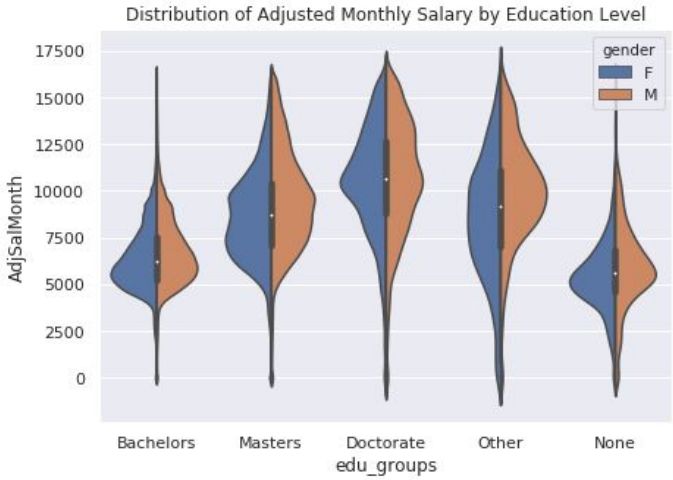


Exploratory Analysis of the Data:

One of our first analyses of teacher salaries was to look at the correlations of the independent variables. The only very strong correlation in our data set (either positive or negative) was between District Level Years of Experience and the State Level Years of Experience, which made sense intuitively. This is something we made sure to keep an eye on for future multicollinearity issues.

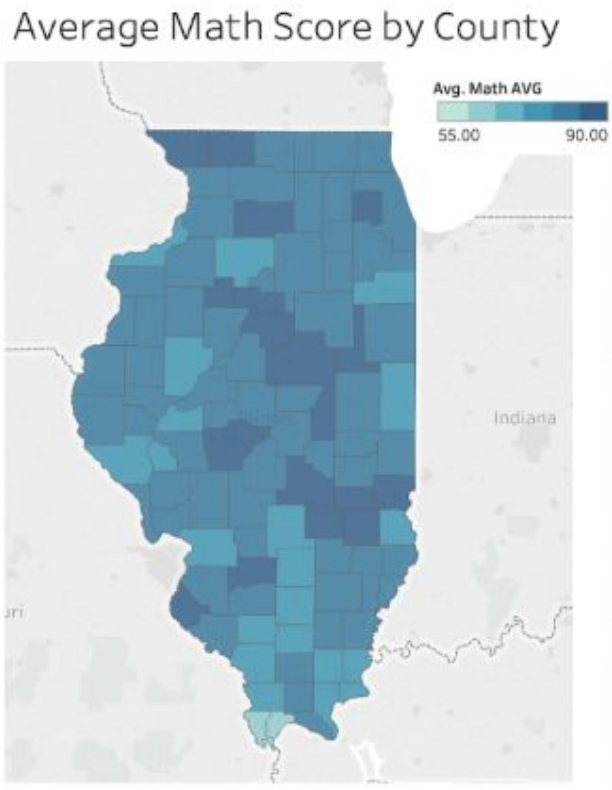
We also made violin plots showing the distributions of Inflation Adjusted Monthly Salaries by various demographic groups as seen below:

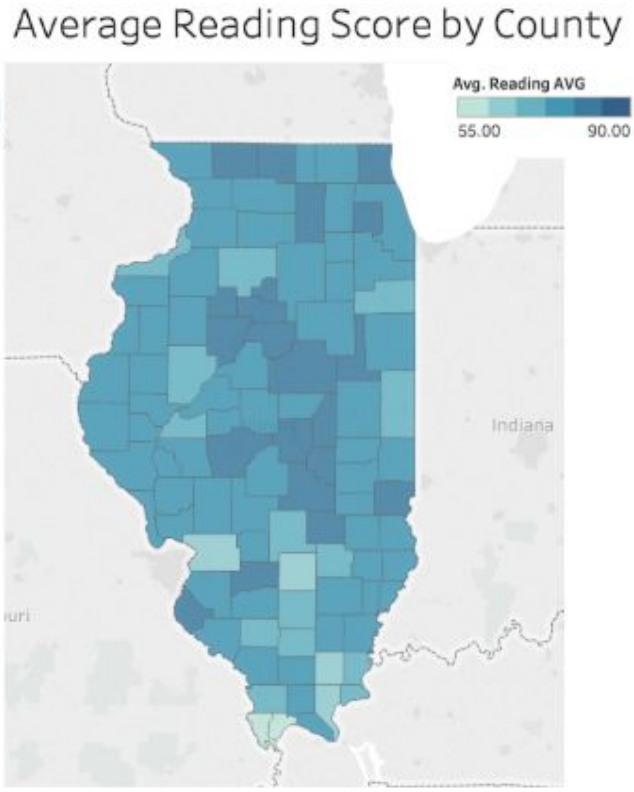


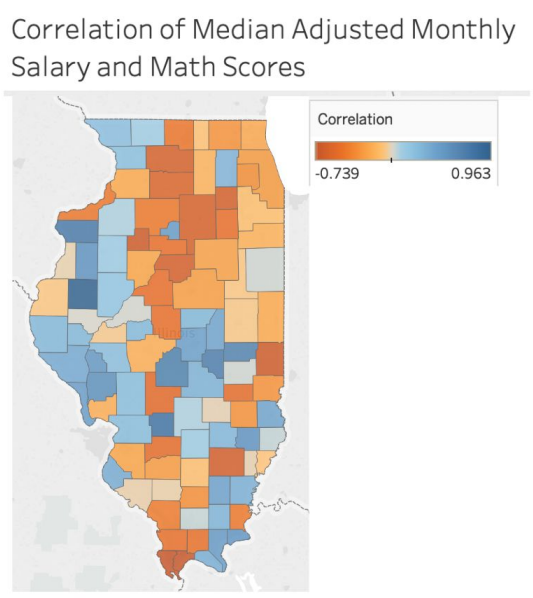


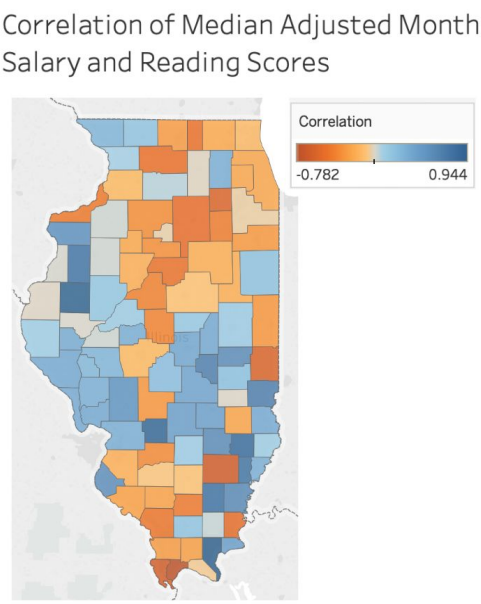
We noticed that at first glance, the median salary for men appears to be higher than for women. We also observed that teachers with a doctorate degree made more than those with a Masters degree, and those with a Masters degree earned more than those with a Bachelors degree. Finally, we saw that the median salary was similar for all ethnic groups except black, which was slightly higher and left-skewed.

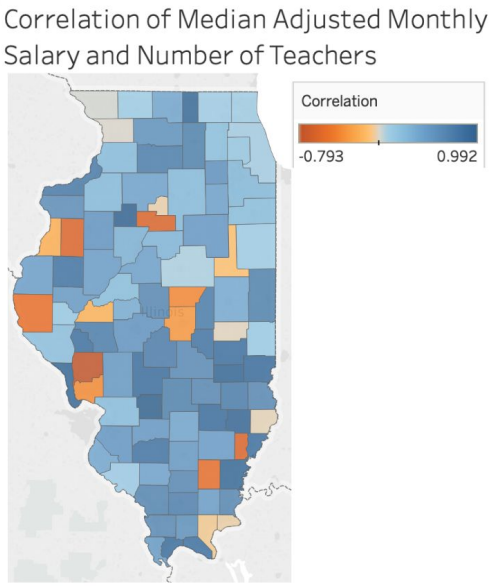
We also analyzed the district level math and reading test scores as seen below:











We observed that the math scores for the district were generally higher than the reading scores and there was a correlation between the scores for each subject. We also found there was a slightly positive correlation between the number of teachers employed and the median monthly salary for each district. Finally, we observed a stronger negative correlation between the test scores for a district and that district's median monthly salary compared to other districts within a county.

Methods Used:

Collaboration Method:

Our group used a relatively new product from Google called Google Colaboratory (or Colab). This is a virtual computing environment (Jupyter Notebook) that allows easy sharing similar to Google Drive. We found this product to be useful for version control as we were all able to be working from the same version of the files at all times. The only drawback was some system crashing and slower performance for models or visualizations since our data set had more than a million rows of data.

Model Methods:

Our initial exploration of potential models included Linear Regression, Decision Tree, Random Forest, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Gradient Boosting, and CatBoost. CatBoost is a new open-source machine learning algorithm that is catered to categorical data (Ray, 2017). It has been shown to be more accurate and faster than algorithms that just create 0/1 dummies for text data. CatBoost can create combinations of categorical features as well as combining categorical and numeric features for more accurate processing. As there were about 1.3 million records, KNN took hours to train and was severely overfitting, so we decided that it was not a feasible model for our analysis. SVMs with various kernel functions all had an explained variance of almost zero. Because of this, the results would not be very meaningful and we were able to safely discard it as a potential model. Baseline models for Linear Regression, Decision Tree, Random Forest, Gradient Boosting, and CatBoost were evaluated and compared using R^2 and RMSE values. We found that Decision Trees performed significantly lower than the other ensemble tree based methods. Although Linear Regression performed well and was not overfitting, we decided to move forward with only Random Forest, Gradient Boosting, and CatBoost because they could better capture and explain the non-linear relationships among the

independent variables and Adjusted Monthly Salary. Additionally, these three models were chosen because tree-based methods inherently produce a list of feature importance that could help explain what features had the biggest influence on predicting teacher salaries. Finally, the tree-based models all had the highest R^2 and lowest RMSE values on our test data set. Grid Search was then performed to find the optimal values for some of the most important hyperparameters that impact model performance. The optimal values determined from grid search were used to evaluate the models moving forward. Then, we performed feature selection using an embedded feature selection approach. Since all three of the models produce a list of the most important features, we compared the model performance using various numbers of the top N features. The goal is to find a more parsimonious model that could be easily interpreted using a smaller feature set and without sacrificing too much in predictive performance.

Feature Selection Methods:

As the objective of our analysis is to explore the factors of teacher salary, our group determined the need to conduct a more in-depth feature selection process. The specifications of our specific problem included a data set with over 1.3 million instances and 40 features. With our resources, we had the ability to conduct a smaller scale analysis of the features using all three types of selection: filter, wrapper, and embedded techniques. For each of the following methods, we used 300,000 random instances from the original dataset.

The first exploratory step was to push the dataset through a low variance filter, which calculates each column variance and removes the columns with a variance lower than 20%. There is no target variable in this filter step, so we are merely analyzing the features importance to each other. We extracted the eight most correlated features from the dataset (see the table below).

The next step was a feature selection method that considered the target variable, Adjusted Monthly Salary. For this, we used two methods: Stepwise Backwards Recursion with Random Forests and a Wrapper Select with a Random Forest Regressor. From these two selection techniques, we determined the features that were the most correlated to our predictor (see the table below).

From all three techniques, the most important features (highlighted in yellow below) are full-time equivalency, district experience, state experience, and first 3 digits of the zip code.

Low Variance	**Stepwise Backwards Recursion - Random Forests**	**Wrapper Select - Random Forest Regressor**
'months_employed' 'fte' 'dist_exp' 'state_exp' 'out_of_state_exp' 'pct_admin' 'first_3' 'grade_taught'	'fte' 'dist_exp' 'state_exp' 'first_3' 'Bachelors'	'months_employed' 'fte' 'dist_exp' 'state_exp' 'first_3' 'grade_taught' 'Bachelors'

Each of these feature sets were then tested with the models to determine which would be the most effective in creating predictive models but not sacrifice predictive performance. For this step, we used Linear Regression, Random Forests, and Gradient Boosting and assessed the R^2 and RMSE values. The table below has the full breakdown of our results.

	Stepwise Recursive Backwards Selection (5 features selected)	*Wrapper Select (7 features selected)*	*Low Variance Filter (8 features selected)*
Linear Regression	Train R^2 - 0.5088 Test R^2- 0.5198 RMSE - 1853.77	Train R^2 - 0.5573 Test R^2- 0.5649 RMSE - 1759.91	Train R^2 - 0.4799 Test R^2- 0.4852 RMSE - 1907.61
Random Forest	Train R^2 - 0.5539 Test R^2- 0.5578 RMSE - 1745.26	Train R^2 - 0.5981 Test R^2- 0.5949 RMSE - 1670.4	Train R^2 - 0.5355 Test R^2- 0.5331 RMSE - 1793.25
Gradient Boosting	Train R^2 - 0.658 Test R^2- 0.6561 RMSE - 1539.13	Train R^2 - 0.7334 Test R^2- 0.7239 RMSE - 1378.95	Train R^2 - 0.6512 Test R^2- 0.6410 RMSE - 1572.46

The feature set that performed the best was from the Wrapper Select with the Random Forest Regressor. There are seven features total in this set: months employed, full-time equivalency, district experience, state experience, the first 3 digits of the zip code, grade taught, and whether a teacher has a Bachelor's degree. The embedded feature selection techniques that were used will be discussed more thoroughly in the specific method section of this paper.

We also wanted to explore the difference between those teachers who taught in primary schools (up to 8th grade) and those who taught in secondary school (9th-12th grade). We split our data and ended up with 884,648 records for primary teachers and 416,921 records for secondary teachers. We ran the same feature selection methods for both data sets independently and they returned the same 5 to 7 features for all feature selection methods. We then compared three models across the three data sets (Total Data, Primary Teachers, Secondary Teachers) and the results are in the table below.

	All Teacher Types (no FS)	All Teacher Types (with FS)	Primary Only (no FS)	Secondary Only (no FS)	Primary Only (with FS)	Secondary Only (with FS)
Random Forest	Train R^2 - 0.641 Test R^2- 0.635	Train R^2 - 0.641 Test R^2- 0.636	Train R^2 - 0.647 Test R^2- 0.640	Train R^2 - 0.679 Test R^2- 0.670	Train R^2 - 0.648 Test R^2- 0.644	Train R^2 - 0.662 Test R^2- 0.666
Gradient Boosting	Train R^2 - 0.785 Test R^2- 0.736	Train R^2 - 0.733 Test R^2- 0.724	Train R^2 - 0.746 Test R^2- 0.734	Train R^2 - 0.752 Test R^2- 0.743	Train R^2 -.738 Test R^2-.731	Train R^2 -.738 Test R^2-.736
CatBoost	Train EV - 0.789 Test EV- 0.787 RMSE - 1126.18	Train EV - 0.760 Test EV- 0.758 RMSE - 1200.59	Train EV - 0.772 Test EV- 0.763 RMSE - 1168.18	Train EV - 0.776 Test EV- 0.768 RMSE - 1435.36	Train EV - 0.745 Test EV- 0.736 RMSE - 1233.02	Train EV - 0.736 Test EV- 0.733 RMSE - 1539.73

With the above results, we decided to not continue along the path of separating the primary and secondary teachers as the features selected were very similar and there was not a significant increase in performance when the data was separated. We were also happy to see that overfitting did not seem to be an issue for us

as the Training and Testing results were similar. This analysis helped us narrow down our choice of algorithms down to two, Gradient Boosting and CatBoost. After more discussion and research we settled on CatBoost as our preferred model going forward.

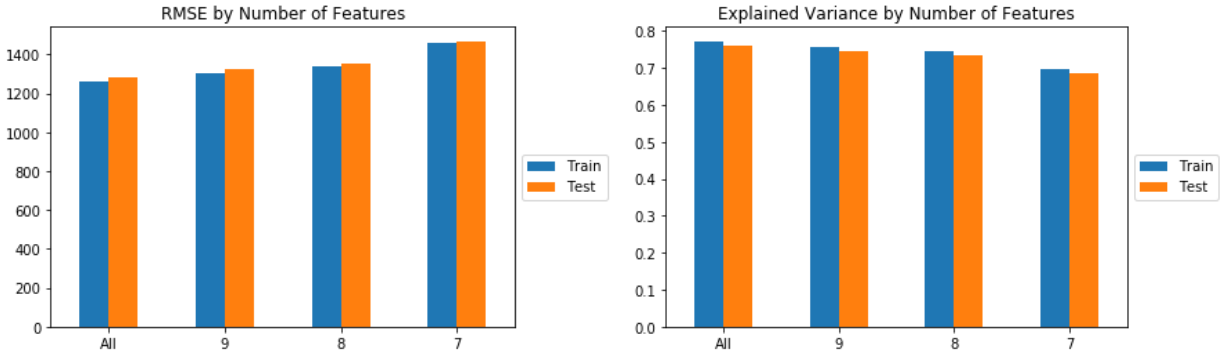
Analysis of Results:

Teacher Salary Dataset:

As KNN and SVM were computationally too expensive and deemed unsuitable for our analysis, we analyzed Linear Regression, Decision Tree, Random Forest, Gradient Boosting, and CatBoost machine learning techniques using the default parameters to evaluate the baseline performance of these models. The results can be found in the Appendix. Decision Tree had the lowest performance with a Training R^2 of 0.533, a Test R^2 of 0.530, and a Test RMSE of 1133.529. Random Forest had a Training R^2 of 0.657, a Test R^2 of 0.656, and a Test RMSE of 1168.438. Linear Regression also performed similarly with a Training R^2 of 0.677, a Test R^2 of 0.677, and a Test RMSE of 1133.529. Gradient Boosting had a Training R^2 of 0.785, a Test R^2 of 0.784, and a Test RMSE of 926.162. CatBoost had the highest performance with a Training R^2 of 0.789, a Test R^2 of 0.787, and a Test RMSE of 1126.18. None of the models appeared to have issues with overfitting. Based on the baseline performance of these models, we decided to focus only on Random Forest, CatBoost, and Gradient Boosting moving forward. CatBoost had the highest R^2 values, while Gradient Boosting had the lowest RMSE. We chose to further investigate Random Forest instead of Linear Regression because there are more opportunities to fine tune the hyperparameters and try to boost model performance.

Grid Search was not possible on the teacher salary dataset due to time constraints and limited computational ability. Instead we had to explore parameter optimization using a more greedy approach based on research and what we found successful for the Test Scores dataset. For Random Forest, we found that the optimal hyperparameters were n_estimators = 50, criterion = ‘mse’, max_depth = 5, min_samples_split = 2, and max_features = 0.33. Using the results from parameter tuning and feature selection with seven features, the Training and Test R^2 values remained about the same at 0.641 and 0.636 respectively. For Gradient Boosting, we found that the optimal parameters were loss = ‘ls’, n_estimators = 300, min_samples_split = 10, and max_depth = 3. Applying these parameters and the seven selected features, Gradient Boosting had a Training R^2 of 0.733 and a Test R^2 of 0.724. For CatBoost we determined that the optimal parameters were n_estimators = 250 and learning_rate = 0.1. Using the seven features that were selected from our feature selection analysis and the optimal parameters, our CatBoost model had a Training R^2 of 0.760, a Test R^2 of 0.758, and a Test RMSE of 1200.59. CatBoost yielded the highest predictive performance for predicting adjusted monthly salaries and was not overfitting, so we decided to choose CatBoost for our final model.

Once we decided on CatBoost, we wanted to do a more specific feature selection analysis. We ran the CatBoost model with All Features, Top 9, Top 8, and Top 7. Results are below:



We noticed a drop in performance when we reduced the number of features to seven so we finalized our feature selection to the top eight features. Again, we did not see a significant difference in the scores for the Test set vs. the Training set so we were not concerned with overfitting our final model. We also compared the relative importance of all features (Appendix). The most important features were zip type (Chicago, Suburb, other), state experience, district experience, months employed, education groups, lowest grade taught, employment description, and the first three digits of the zipcode. We were not too concerned with multicollinearity in our model as the only highly correlated features were District experience and State experience, which were both found to be important in the final model. CatBoost is based on Decision Trees which themselves are less susceptible to multicollinearity as they will only select one of the correlated features at a time for a split. With this, we were not concerned about the correlation among our final feature set. Our final model with eight features had a Training R^2 of 0.745, a Test R^2 of 0.734, and a Test RMSE of 1353.02.

Test Scores Dataset:

We analyzed the Test Scores dataset with seven machine learning techniques using the default parameters to evaluate the baseline model performance. The results are displayed in the Appendix. SVM with a linear kernel function performed the lowest with a Training R^2 of 0.054, a Test R^2 of 0.052, and a Test RMSE of 1264.420. Linear Regression also performed poorly with a Training R^2 of 0.264, a Test R^2 of 0.278, and a Test RMSE of 1103.366. Decision Tree performed slightly better with a Training R^2 of 0.341, a Test R^2 of 0.315, and a Test RMSE of 1075.235. KNN had better scores with a Training R^2 of 0.453, a Test R^2 of 0.396, and a Test RMSE of 1009.703, however, the Training R^2 was significantly higher than the Test R^2, indicating that there are some issues with overfitting. Random Forest had a Training R^2 of 0.446, a Test R^2 of 0.409, and a Test RMSE of 998.672. CatBoost had a Training R^2 of 0.446, a Test R^2 of 0.442, and a Test RMSE of 1045.184. Gradient Boosting had the highest performance with a Training R^2 of 0.577, a Test R^2 of 0.492, and a Test RMSE of 925.760. Based on the baseline performance of these models, we decided to focus only on Random Forest, CatBoost, and Gradient Boosting moving forward. Although Gradient Boosting had the highest R^2 scores and lowest RMSE, we have to monitor potential overfitting. CatBoost had a slightly higher RMSE than the other two modeling techniques. However, the Training and Test R^2 values were very similar which may indicate that CatBoost is not overfitting and may generalize better.

Next, we utilized Grid Search to determine the optimal values for some of the most important hyperparameters that impact model performance. For Random Forest we found that the optimal values were n_estimators = 1000, criterion = ‘mse’, min_samples_split = 0.001, min_samples_leaf = 1, and max_depth = 5. After training the Random Forest applying the optimal parameters from Grid Search the Training R^2 increased to 0.508, the Test R^2 increased to 0.454, and the Test RMSE decreased to 973.166.

For Gradient Boosting we found that the optimal values were n_estimators = 500, loss = 'ls', learning_rate = 0.01, criterion = 'friedman_mse', min_samples_split = 0.001, and max_depth = 4. Applying the optimal parameters from Grid Search, the Gradient Boosting Training R^2 increased to 0.608, the Test R^2 increased to 0.516, and the Test RMSE decreased to 915.954. For CatBoost we found that the optimal parameters were n_estimators = 500, loss_function = 'RMSE', and learning_rate = 0.1. After applying the optimal parameters from Grid Search to CatBoost, the Training R^2 increased significantly to 0.800, the Test R^2 increased to 0.730, and the Test RMSE decreased to 685.203. We found that the number of estimators played a large role in increasing predictive performance of our models, especially for CatBoost. With 100 estimators the location features were the least important features in predicting teacher salaries, however, when the number of estimators was increased to 500, Region and County were the most important features in predicting teacher salaries and the Test R^2 increased to from 0.442 to 0.730. Since there are 102 counties in Illinois, it may have taken a long time to learn the complex relationship between County and the adjusted monthly salary, where 100 estimators was not enough.

Since CatBoost had the highest R^2 and lowest RMSE, we decided to choose CatBoost for our final model for the Test Scores dataset. The order of feature importance was Region, County, Number of Teachers, Math AVG, Enrollment, Reading AVG, Students/Teacher, and District Number. We then performed an embedded feature selection using the results from the list of feature importance to see if we could get a more parsimonious model with a similar predictive performance. However, removing only the least important feature decreased the Training R^2 from 0.800 to 0.707, the Test R^2 from 0.730 to 0.623, and increased the Test RMSE from 685.203 to 805.490. This is a significant change in predictive performance, so we decided to keep all eight of the original features in the final model. Even though we could not further reduce the feature set, eight features is not that many and could still be easily interpreted and explained.

Validation and Testing of Final Model:

All models were compared using the same train/test/validation split of the dataset to ensure that differences in performance could be attributed to the modeling techniques themselves and not the random split of the dataset. We also used the same random state for individual models when fine tuning the hyperparameters. Once we chose CatBoost for our final model, we validated model performance by comparing the training results to unseen test and validation datasets. The test and validation results were similar to the training results, indicating that there were not issues with the model overfitting and that the model should generalize well to new data. The residual plots of the final model for the test and validation sets can be found in the Appendix. These plots demonstrate that our final model typically underpredicts the adjusted monthly salaries. Finally, we validated our final model performance by comparing our results to other train/test/validation splits and changing the random state of the CatBoost Regressor. This verified that our results were not an anomaly and were representative of overall model performance. This was especially important for the Test Scores dataset because we observed such a large boost in model performance when optimizing the CatBoost parameters.

Discussions and Conclusions:

After conducting our analysis, the results disproved our hypothesis. While the original data set affirmed our bias and showed evidence of salary bias, especially among gender and ethnicity, the model shows a different picture. This demographic information, such as gender and race/ethnicity, are not identified as the 'most important' predictors for teacher salary in our final CatBoost model. This means that while a

wage gap still exists based on choice demographics, these are not features that determine teacher salary. The main factors that contribute to the model are related to education/experience level and location. These are the features that show a relatively high level of wage inequality. Of the eight selected features in the final CatBoost model, Zip type, Education Groups, First 3 digits of Zip Code, and even State and District experience are all related to education and location.

In addition, the CatBoost model on the Test scores data set disproved the current push for performance-based salary. Should teacher salaries depend on the performance of their students on standardized tests? From our analysis, there is no strong correlation between higher test scores and higher salaries. However, there is a positive correlation between the number of teachers employed in a district and higher salaries. The possibilities for this correlation could mean that teachers have more bargaining power in denser districts or that more teachers mean more value is placed on education and thus, the district is willing to pay teachers more.

The importance of these features in teacher salaries leads to the assumption that there is a much larger systematic problem causing this wage inequality. Factors such as school funding and additional benefits such as pensions, could have a bigger impact on the inequalities in teacher compensation that should be investigated more thoroughly.

Future Work:

The conclusion of the analysis leads to questions about how teacher salary compares to different factors across the state of Illinois. With more time, the incorporation of a more comprehensive analysis of how test scores affect teacher salary would benefit districts and the state to make educated decisions on whether or not this should be a factor for pay.

As there are wage inequalities between different education levels and discrepancies based on location, further analysis would benefit from a more granular analysis of teacher salaries that eliminate these factors. The next step of the analysis would be to look at teacher salaries across schools within the same district - thus, eliminating the location factor. To eliminate the education wage gap, the analysis would require the separation of different levels, such as analyzing those with Bachelor degrees, then those with Master degrees, and so on. Taking these steps would allow a different sample of features into the creation of the model and could lead to us seeing other, potentially demographic, factors as predictive indicators in teacher salary.

In addition, the current state of the teacher salary discussion revolves around investigating the growing wage gap between teachers and other professional careers. This could have an impact on salaries within the teaching community so looking at the bigger professional picture might be a good step in seeing how teacher salaries compare to their equivalent counterparts.

References:

Allegretto, Sylvia, and Lawrence Mishel. “The Teacher Pay Penalty Has Hit a New High: Trends in the Teacher Wage and Compensation Gaps through 2017.” *Economic Policy Institute*, 2018, www.epi.org/publication/teacher-pay-gap-2018/.

Hendricks, Matthew D. “Towards an Optimal Teacher Salary Schedule: Designing Base Salary to Attract and Retain Effective Teachers.” *SSRN Electronic Journal*, 2013.

Huang, Guomin, et al., “Evaluation of CatBoost Method for Prediction of Reference Evapotranspiration in Humid Regions.” *Journal of Hydrology*, vol. 574, 29 Apr. 2019, pp. 1029–1041.

Khongchai, Pornthep, and Pokpong Songmuang. “Random Forest for Salary Prediction System to Improve Students' Motivation.” *2016 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, 2016.

Ray, Sunil. “CatBoost: Machine Learning Library to Handle Categorical Data Automatically.” *Analytics Vidhya*, 4 Sept. 2017, www.analyticsvidhya.com/blog/2017/08/catboost-automated-categorical-data/.

Appendix:

Description of Final Code Files:

- Capstone_Teacher_Salary.ipynb
    - Loads Data, Creates .csv files to be used in other files, Data Preprocessing, Outlier and Exploratory Analysis, First Attempts at Models without Feature Selection
- TeacherSalaryModels.ipynb
    - More detailed analyses with Feature Selection for Teacher Salary Data only
- TestScoresModels.ipynb
    - More detailed analyses with model development and Feature Selection for Test Score Data only
- PrimaryTeacherSalaryModels.ipynb
    - Model development for Primary Teachers Only
- SecondaryTeacherSalaryModels.ipynb
    - Model development for Primary Teachers Only
- CatBoost.ipynb
    - Final model chosen with all analysis and performance metrics

Teacher Salary Data Set Baseline Model Performance:

	Train R^2	Test R^2	Test RMSE
Linear Regression	0.677	0.677	1133.529
Decision Tree	0.533	0.530	1362.662
Random Forest	0.657	0.656	1168.438
Gradient Boosting	0.785	0.784	926.162
Catboost	0.789	0.787	1126.18

Final Catboost Model for Teacher Salary Data Set:

	No FS	9 Features	8 Features	7 Features
Train RMSE	1262.14	1303.57	1335.71	1458.67
Test RMSE	1283.93	1324.06	1353.02	1468.65
Train R^2	0.7723	0.7571	0.7449	0.6958
Test R^2	0.7606	0.7454	0.7342	0.6868

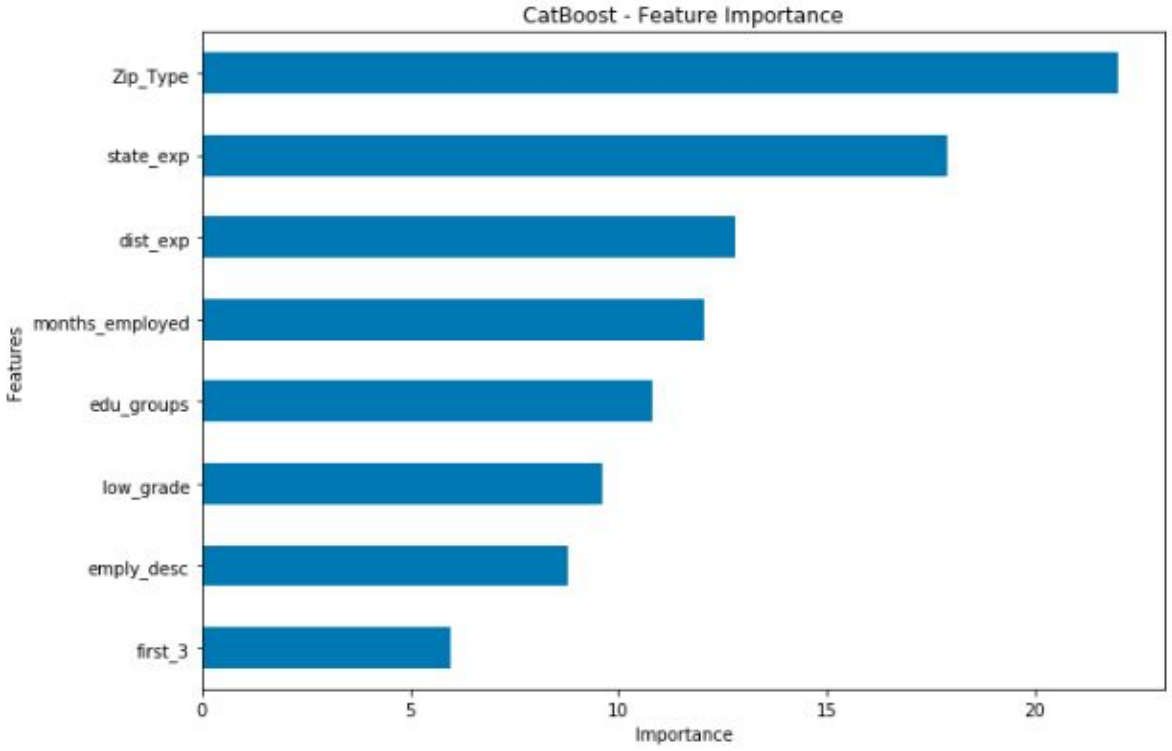
Test Scores Data Set Baseline Model Performance:

	Train R^2	Test R^2	RMSE
Linear Regression	0.264	0.278	1103.366
Decision Tree	0.341	0.315	1075.235
SVM w/ linear kernel	0.054	0.052	1264.420
Random Forest	0.446	0.409	998.672
KNN	0.453	0.396	1009.703
Gradient Boosting	0.577	0.492	925.760
CatBoost	0.446	0.442	1045.184

Final Catboost Model for Test Scores Data Set:

	Baseline Model	Grid Search	Feature Selection
CatBoost	Train R^2 - 0.446 Test R^2- 0.442 RMSE - 1045.184	Train R^2 - 0.800 Test R^2- 0.730 RMSE - 685.203	Train R^2 - 0.707 Test R^2- 0.626 RMSE - 805.490 7/8 features

Feature Importance from CatBoost model:



Final Catboost Model Plots showing Predicted vs. Actual: