Estimating Income Changes

The Why

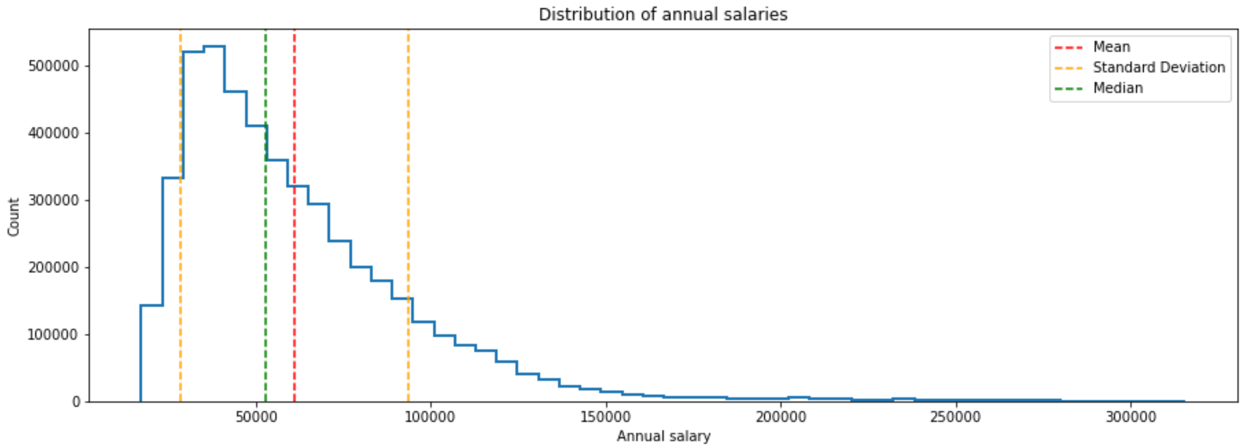
Future of work is changing very quickly. More people have the option to work remotely and many are learning new skills in hope of better jobs or higher income. New technology can sometimes require certain new skills becoming more important in a particular occupation. We want to be able to estimate this new income and the amount of change from the original salary estimate.

The Data

The data for annual income and employment was obtained from the Bureau of Labor Statistics (BLS) and includes salaries of 808 unique occupations from all 50 states, District of Columbia, and Puerto Rico. Skills data came from the O*net website <describe website here> and identifies 35 unique skills and two different ways of rating them, “importance” and “level”. These each have their own scale. In this report we describe a model based only on “importance” values, which has a scale between 1 and 5, inclusive. Future work integrating the “level” scale should be considered in order to refine the model.

What we are trying to predict

Our dependent variable or target feature is “annual income”. BLS data includes mean, median, and the 10th, 25th, 75th and 90 percentiles on both “annual” and “hourly” incomes. In most cases, annual income was calculated by multiplying the hourly value by 2,080 hours, the full-time, year-round value. There are occupations where no hourly data was given but an estimate of annual income was still reported. These jobs include teachers, sportsmen and their agents and <insert more examples here>. The reverse is also true for such jobs as gig workers and actors where an hourly income is available but they do not work on a 2,080 hours per year standard which makes it difficult to estimate annual salaries. Between the two options annual and hourly salaries as the values we would like to estimate, since annual is missing fewer values (2 percent compared to 6.6 percent for hourly) than hourly, giving us more data values to work with. After choosing annual income as our target feature, we also dropped any rows with missing values in that column.



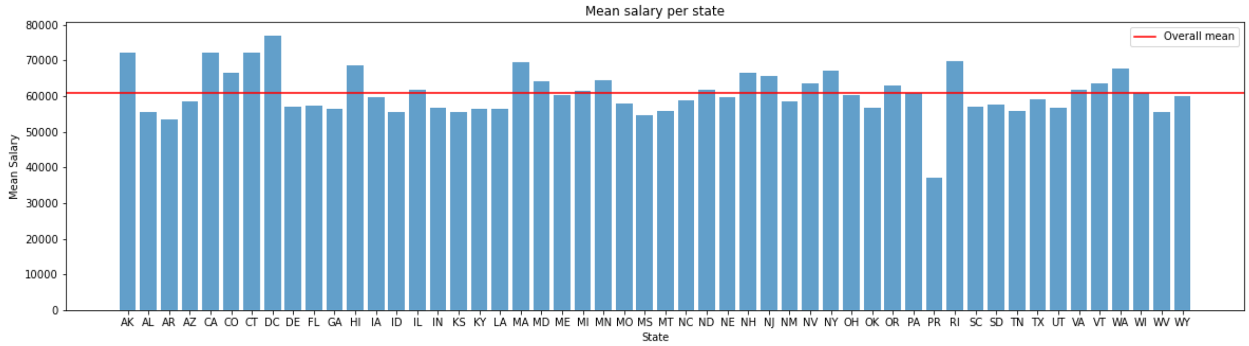
Features selection

The O*net data on skills uses the same standard as BLS for occupation codes, so we were easily able to combine the two datasets. In some cases O*net data had more specific job titles, such as Baristas vs Fast Food and Counter Workers or Financial Managers for both Investment Fund Managers and Treasures and Controllers. After verifying accurate associations, we replace the more general occupation titles of BLS, with the O*net titles. There are 35 unique skills identified in our data-set. <Give examples?>. Each occupation is ranked using two different continuous scales, importance (1-5) and level (0-7). For this project we worked with the importance scale.

Our final data-set has 4,774,000 rows corresponding to 788 unique occupations (down from 808 jobs) and six columns which are occupation name, skill, importance value, location, state, and annual salary.

There are 379 unique cities included. Instead of working with these many added features, we will use the average statewide income for each occupation. In this manner we will end up with 52 features (after one-hot encoding the states), instead of 379.

In order to determine if states should be treated equally or not, we compare the average income of each state with the overall mean.



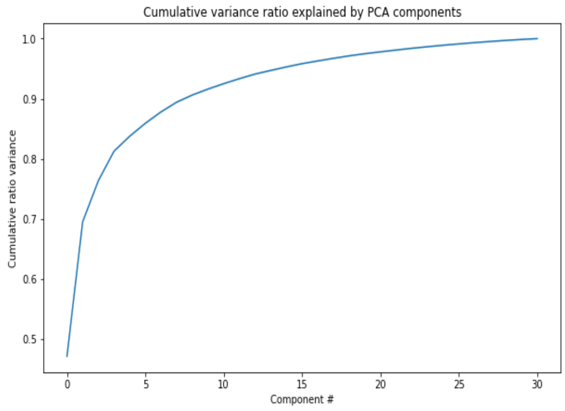
From the plot we can see that living in Puerto Rico may give someone a large disadvantage whereas moving to California or DC for example might result in an increase in income. So states will remain as features, which gives us 52 additional features to work with.

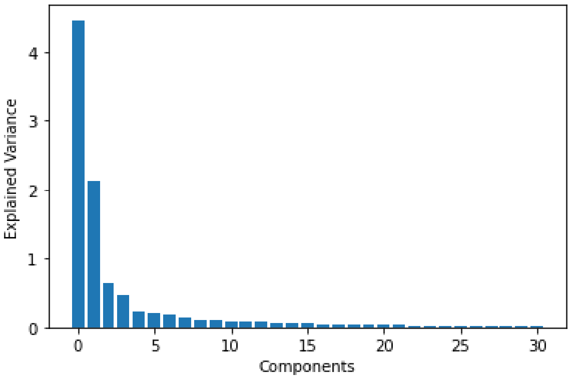
Looking for the right dimensions

During EDA, from distribution of skill values, the four skills Installation, Equipment Maintenance, Equipment Selection and Repairing are highly skewed to the left. These skills don’t seem to be very important for the majority of jobs out there and most of the values sit at the minimum importance value of 1. So I have eliminated these four features leaving us with 31 skills and 52 states, giving 83 features total.

Looking at the correlations between features, there are some features that are highly correlated such as Systems Analysis and Systems Evaluations for example.

Running a PCA analysis allows us to estimate the best dimension for our problem. With 83 features, we risk overfitting.





Approximately 85 percent of the variance is explained by the first 5 PCA components and close to 90 percent by the first 8 components.

Preprocessing

Our final data set has 83 features and 136,041 observations representing average numbers for a particular occupation in a particular state. This data was then split into a train/test set with the 30 percent of the data set aside for testing.

For our preprocessing, in addition to one-hot encoding state labels and splitting the data into a train and test split, we should also standardize the data since skill measures are on a continuous scale from one to seven while state values are binary. For this reason, I have chosen to standardize my data, using minmax method in most cases. <insert exception here>

Modeling

We ran three Regression and one Random Forest Model. The results are summarized below:

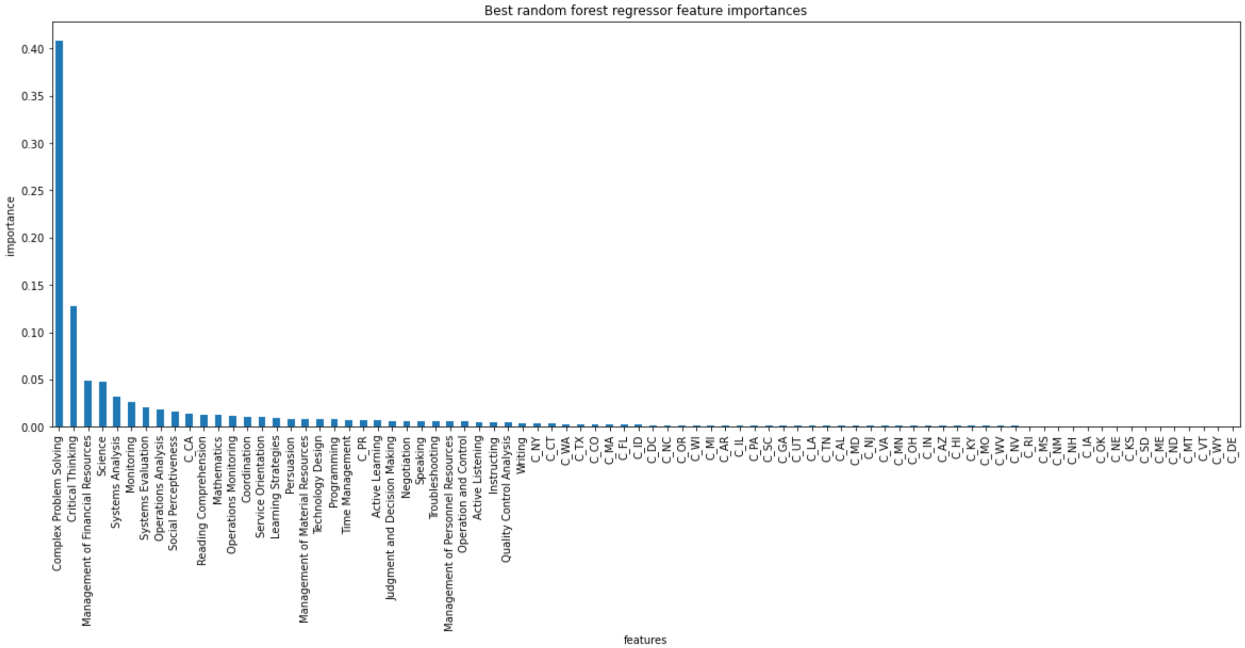
	OLS	Ridge	SVR	Random Forest
R-squared **(variance)**	0.61	0.61	0.59	0.89
Mean Absolute Error	$14,251	$14,250	$13,671	$6,995

Results - Best Model

Our best model is the Random Forest Model with 1000 estimates and a mean absolute error of approximately $7,000. The top 10 most important features (skills) are very similar for all models. Management of financial resources and science are in the top five most important skills in all models. <which ones are available in all>. Note that living in California does make the top 10 in importance.

Random Forest Model feature importance

1. Complex Problem Solving
2. Critical Thinking
3. Management of Financial Resources
4. Science
5. Systems Analysis
6. Monitoring
7. Systems Evaluation
8. Operations Analysis
9. Social Perceptiveness
10. C_CA



Model in action

- Aspects of a position change, for example a certain skill becomes more important for a job, or two positions merge and more skills are necessary to have for someone in that position. How much should the current salaries increase if at all?

- A remote employee moves to a more expensive state. Are they entitled to more money? What if they move to a less expensive state? Should their current income be re-evaluated?
- When offering paid learning to employees, which skills provide the most value to the position while save the company money by costing the least due to resulting salary increases?

Future work

- Incorporate the level scale into the model
- Reduce dimensionality using the most common top skills based on PCA results. Or by clustering, number of clusters being determined by a particular situation.
- Try more models and hyperparameters
- Multiple skill/move change