Final Report

Cuthbert Chow, Rong Li, Andy Yang

2021-11-27

Contents

Aim and Summary

The main predictive question we wish to answer is what we can expect a person's salary to be in the US, given a certain professional history (such as years of experience, industry, or age). We will use a linear regression model to do the prediction. In the process, we wish to understand which factors provide the most predictive power when trying to predict a person's salary.

Methods

The Python (Van Rossum and Drake 2009) and R (R Core Team 2021) programming languages and the following Python and R packages were used to perform the data analysis and present results: Pandas (Reback et al. 2020), Scikit-learn (Pedregosa et al. 2011), Altair (VanderPlas et al. 2018), docopt (Keleshev 2014), knitr (Xie 2021).

As references, we utilized (Jain 2017) for methodological practices regarding linear, ridge and lasso regression, as well as (Martín et al. 2018) which recommended linear regression for problems similar to the one we are analysing.

Results & Discussion

Data

The dataset we are analysing comes from a salary survey from the "Ask a Manager" blog by Alison Green. This dataset contains survey data gathered from "Ask a Manager" readers working in a variety of industries. (Green 2021)

We used Altair (VanderPlas et al. 2018) to create figures, Pandas (Reback et al. 2020) to do data processing, and Scikit-learn (Pedregosa et al. 2011) to perform statistical analysis.

Results and Discussion

First, we looked at the distribution of our target "Annual Salary." As shown in the graph below, it seems to be a largely right-skewed distribution. And the median salary is around $80,000.

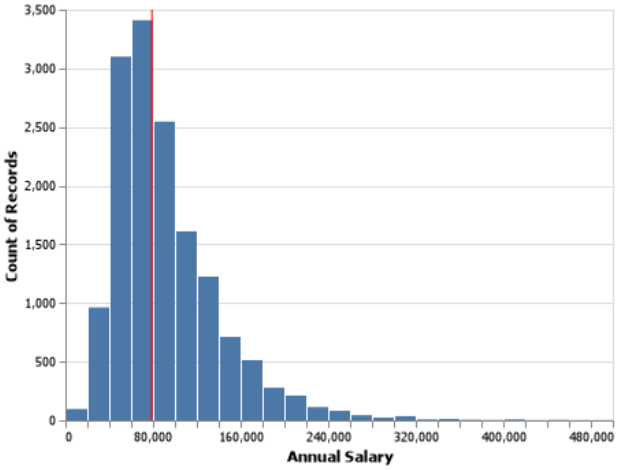

Figure 1: Figure 1 - Distribution of Annual Salaries

Here is some general information about our dataset:

To look at whether the features in our dataset are useful to predict annual salary, we first looked at a summary table about our features:

Table 1: Table 1 - Summary Information About Key Features

Features	Not.Null.Count	Null.Count	Number.of.Unique.Values	Unique Values	Types
how_old_are_you	15037	0	7	[‘45-54,’ ‘25-34,’ ‘35-44,’ ‘55-64,’ ‘65 or over’]	object
industry	15008	29	675	[‘Accounting, Banking & Finance,’ ‘Engineering or Manufacturing,’ ‘Education (Higher Education),’ ‘Computing or Tech,’ ‘Health care’]	object
job_title	15037	0	7970	[‘CPA,’ ‘Sales Analyst 1,’ ‘Director of Enrollment,’ ‘Process Analyst,’ ‘Senior Data Scientist’]	object
other_monetary_comp	11282	3755	583	[10000.0, 2700.0, 0.0, 5000.0, 145000.0]	float64
state	14914	123	108	[‘California,’ ‘Pennsylvania,’ ‘Colorado,’ ‘Virginia,’ ‘Oregon’]	object
city	15006	31	2482	[‘Palm Springs,’ ‘Pittsburgh,’ ‘Fort Collins,’ ‘Arlington,’ ‘Boulder’]	object
overall_years_of_professional_experience	15037	0	8	[‘21 - 30 years,’ ‘11 - 20 years,’ ‘8 - 10 years,’ ‘2 - 4 years,’ ‘5-7 years’]	object
years_of_experience_in_field	15037	0	8	[‘8 - 10 years,’ ‘5-7 years,’ ‘11 - 20 years,’ ‘2 - 4 years,’ ‘1 year or less’]	object
highest_level_of_education_completed	14935	102	6	[“Master’s degree,” ‘College degree,’ ‘Some college,’ ‘PhD,’ ‘High School’]	object

We noticed that there are lots of null values in the additional information features (additional_context_on_job_title, additional_context_on_income, etc), and some of the variables have a lot of unique values.

Here we want to explore those variables that have < 10 unique values and check their distributions and relationships with the annual salary, since variables with 100s or 1000s of distinct values would be harder to visualize in a meaningful way.

As shown below, the higher salaries are roughly associated with the older age groups, the longer experience and the higher education, which indicates those are likely to be good predictors of our target.

We chose a linear Ridge regression model with the alpha hyperparameter to predict annnual salary based on the given features in the dataset. To ensure that the model was not overfitting to training data, we conducted some additional data cleaning. Firstly, *annual_salary* values within the training dataset of less than 10,000 USD or over 1,000,000 USD were removed. Additionally, text values that occured less than 5 times in the *state* or *city* features were imputed with an empty string. This ensures that highly specific values will be removed which ultimately helps reduce overfitting.

r2	Negative.RMSE	alpha
0.4870195	-38159.66	1.832981e+00
0.4746404	-38617.90	5.455595e-01
0.4709605	-38752.08	6.951928e+01
0.4622520	-39071.07	1.623777e-01
0.4547682	-39342.34	4.832930e-02
0.4509969	-39477.61	1.438450e-02
0.4496944	-39523.69	4.281300e-03
0.4495908	-39527.74	3.793000e-04
0.4495321	-39529.64	1.274300e-03
0.4488644	-39553.48	1.000000e-05
0.4486738	-39560.46	1.129000e-04
0.4486148	-39562.29	3.360000e-05
0.4399704	-39871.13	2.335721e+02
0.3976762	-41349.37	7.847600e+02
0.3402341	-43276.44	2.636651e+03
0.2562765	-45948.20	8.858668e+03
0.1504469	-49107.84	2.976351e+04
0.0652891	-51508.79	1.000000e+05

Using this hyperparameter value, a Ridge model was fitted to the training data and evaluated on the test data. The results can be seen in the table below.

Table 3: Table 3 - Scores of Ridge Model on Test Data

Metric	Scores
R2	0.38
RMSE	48430.31

The results suggest that our model has a hard time accurately predicting the annual salary targets in the test set, with a r2 value of 0.38. This suggests that we may need to further tune our model with feature engineering, or Ridge may not be a good fit for this problem.

To visualize the effectiveness of our model, we can plot the predicted salary values against the actual salary values and compare the correlation to a 45 degree line.

The graph above suggests that the model has high variance and is affected by a large number of outliers within the 50-150 thousand range for predicted salary, which explains the poor performance of the model.

References

Green, Alison. 2021. "How Much Money Do You Make?" *Ask A Manager*. https://www.askamanager.org/2021/04/how-much-money-do-you-make-4.html.

Jain, Shubham. 2017. "A Comprehensive Beginners Guide for Linear, Ridge and Lasso Regression in Python and r." *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2017/06/a-comprehensive-guide-for-linear-ridge-and-lasso-regression/.

Keleshev, Vladimir. 2014. *Docopt: Command-Line Interface Description Language*. https://github.com/docopt/docopt.

Martín, Ignacio, Andrea Mariello, Roberto Battiti, and José Alberto Hernández. 2018. "Salary Prediction in the IT Job Market with Few High-Dimensional Samples: A Spanish Case Study." *International Journal of Computational Intelligence Systems* 11: 1192–1209. https://doi.org/https://doi.org/10.2991/ijcis.11.1.90.

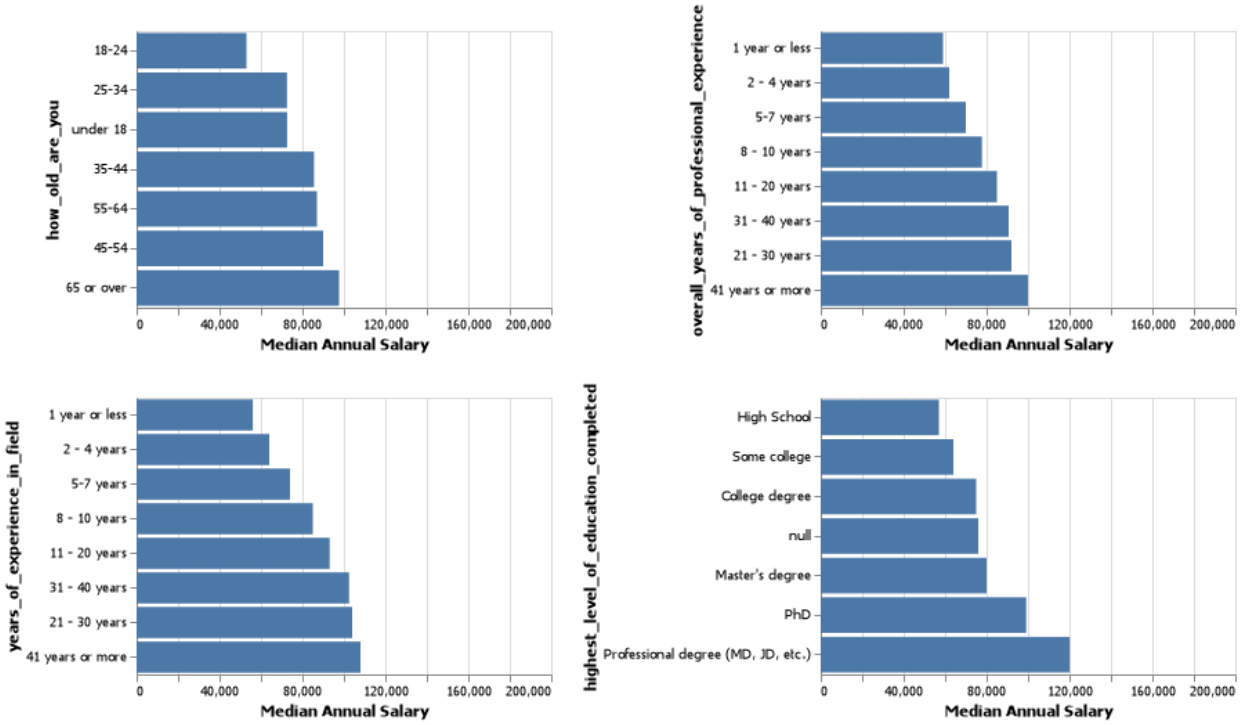

Figure 2: Figure 2 - Median Salary For Various Categorial Features

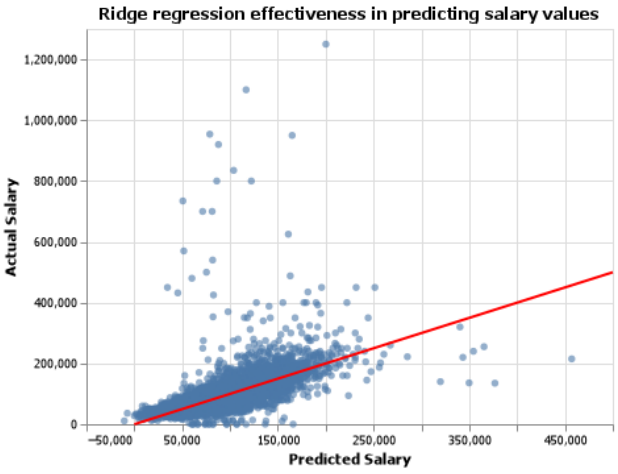

Figure 3: Figure 3 - Actual vs Predicted Salary Values

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12 (Oct): 2825–30.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Reback, Jeff, jbrockmendel, Wes McKinney, Joris Van den Bossche, Tom Augspurger, Phillip Cloud, Simon Hawkins, et al. 2020. *Pandas-Dev/Pandas: Pandas* (version latest). Zenodo. https://doi.org/10.5281/zenodo.3509134.

Van Rossum, Guido, and Fred L. Drake. 2009. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

VanderPlas, Jacob, Brian Granger, Jeffrey Heer, Dominik Moritz, Kanit Wongsuphasawat, Arvind Satyanarayan, Eitan Lees, Ilia Timofeev, Ben Welsh, and Scott Sievert. 2018. "Altair: Interactive Statistical Visualizations for Python." *Journal of Open Source Software* 3 (32): 1057.

Xie, Yihui. 2021. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.