Georgia Institute of Technology
H. Milton Stewart School of Industrial and Systems Engineering

ISYE 6414
Statistical Modeling and Regression Analysis

Final Project – Group 5
Human Resource Data Analysis

Authors:
Te-Kai Chou, Yuebai Gao, Marianus Kick, Manav Sheth

Date of Submission:
December 10th, 2021

Summary

This report shows the application of a linear regression model to a data set in Human Resource Management. The relationships between various predictors and the response variable, salary, are investigated. We found that position and employment satisfaction are some of the major drivers for salary. Through various variable selection techniques, we finally propose the linear regression model chosen by elastic net regularized regression due to its superior overall performance. Neither variables sex or race are chosen as predictors by the elastic net model, supporting the possible conclusion that the salary data does not show a gender or racial bias.

Table of Contents

1. Introduction

In today's world, data-driven decision making is becoming increasingly important for business success. As a result, the field of Human Resource Management is striving to implement analytic tools into their key operations. Employee salaries are critical to companies since they not only directly impact employee productivity but are also a key to attract high-skilled employees. At the same time salaries remain one of the main cost factors for any cooperation. A successful management of associated challenges is therefore essential to companies' performance. What's more, understanding the drivers of salary is important to individual employees to. It can not only help an them to increase personal benefits, but also understand where they stand and help in evaluation processes of their current job position and potential job offers. [1] [2]

1.1. Problem Statement

In this project, our goal is to understand the relationships between employees' salary and predicting variables including work related indications like performance and satisfaction assessments as well as information about personal backgrounds like marriage status. Results could be used by the company to gain insights about main drivers in their current salary structure as well as a first step in refining current salary definition processes. Furthermore, such analysis may answer related questions about gender equality and racial biases which may be relevant for public relationship management. [3] [4]

1.2. Expected Results

Before the analysis, we made several assumptions about how different variables could determine a person's salary. We assume that position and duration since hired would greatly affect one's salary. The performance score, engagement level, employee satisfaction and special projects count might also drive the final salary. Meanwhile, we propose there might be gender and racial biases in terms of salary, in other words, there might be noticeable salary gaps between different genders as well as among different races. Finally, we assume that a regular linear regression model will be sufficient.

1.3. Data

The data set was obtained through online research, provided on Kaggle.com [5]. The data set deals with information related to human resources (HR) of one specific company. This data set was created and is still being maintained by Dr. Carla Patalano and Dr. Richard Hubner from Cambridge College. The data set contains 311 observations on 36 variables. To make the data more digestible, we heuristically reduced the number of variables to 21. This was possible due to replications and clearly unrelated variables. For data cleaning, we first created a new variable, duration since hired (*EmployedYear*), using 2021 minus hiring year. We also corrected a few mistakes: unify entries for *HispanicLatino* to be only 'Yes' or 'No' as well as removing the entry whose race description is Hispanic.

In order to evaluate the performance of the model, we split the data into training data and testing data at a ratio of 7:3. However, we faced the issue that some features of testing data are not in training data when evaluating the model. Therefore, we added the testing data absence in training data into the training model.

2. Analysis

This chapter contains the main implementation of this regression analysis. An exploratory data analysis and an initial linear regression model pave the way towards a thorough model selection process.

2.1. Exploratory Data Analysis

First, we created boxplots and scatterplots between each predicting variable and the response variable, salary. Figure 1 shows that *Position* yield explanatory power for salary, which matches with our assumptions. From the plot, managerial positions like CIO, CEO and Directors have the highest salary when Administrative Assistants have the lowest salary.

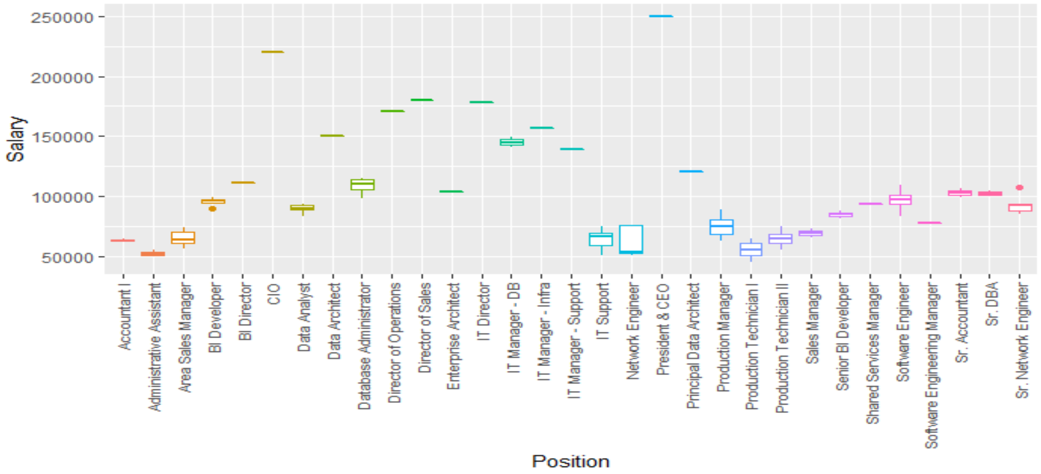


Figure 1: Boxplot, Salary vs. Position

In terms of department, the Executive office again has the highest median salary followed by IT and Software Engineering. The median salary between male and female does not appear to have a great difference when more males are concentrated above median than females and more females have high salary above 175K. Similar situation happens among different races, that is the medians do not have big differences while Asians and Black or African Americans have more entries above median, and more whites have outliers of high salary above 175K. See Figure 16 to Figure 20 in the appendix for associated plots.

Among the scatter plots, Figure 2 shows that the number of special projects accomplished (*SpecialProjectsCount*) seems to have the biggest positive correlation with salary followed by performance score (*PerfScoreID*) shown in Figure 3.

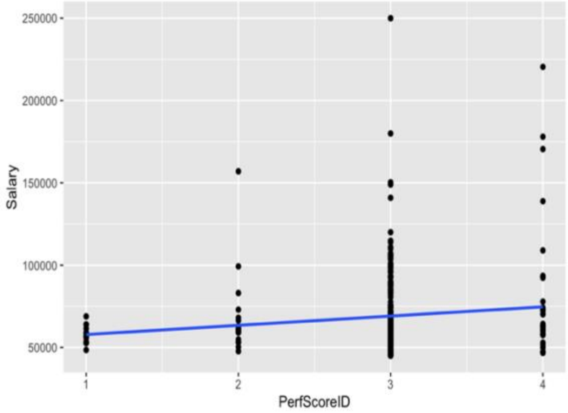


Figure 2: Scatter Plot, Salary vs. PerfScoreID

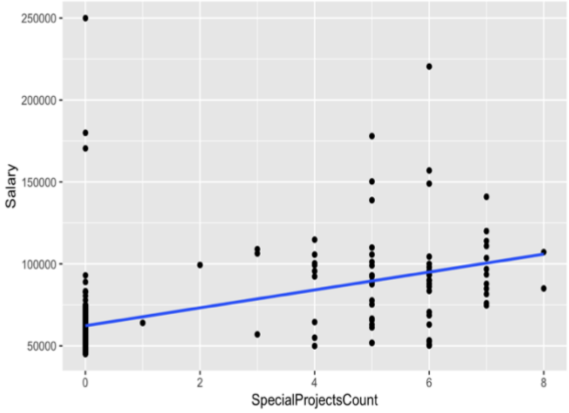


Figure 3: Scatter Plot, Salary vs. SpecialProjectsCount

Surprisingly, duration since hired (*EmployedYear*) does not have too much impact on salary while the age of an employee has slightly bigger positive correlation with salary. Just as we expected, engagement level and employment satisfaction both have a positive correlation with salary. Interestingly, the number of days late in the last 30 days (*DaysLateLast30*) has a negative correlation with salary when the number of absent days (*Absences*) even has a positive correlation with salary. See Figure 21 to Figure 24 in the appendix for additional scatter plots.

Exploratory analysis furthermore shows that there are very few observations for certain categories, such as the *Department* of Executive office, *MaritalDesc* of widowed or employment status of terminated for cause. Such uneven distribution of observations could decrease the prediction accuracy. See Figure 13 to Figure 15 in the appendix for associated plots.

From Figure 4, we would know that observations with the production department take up more than half of the dataset. Within the production department, white is the most occurred race followed by black or African American.

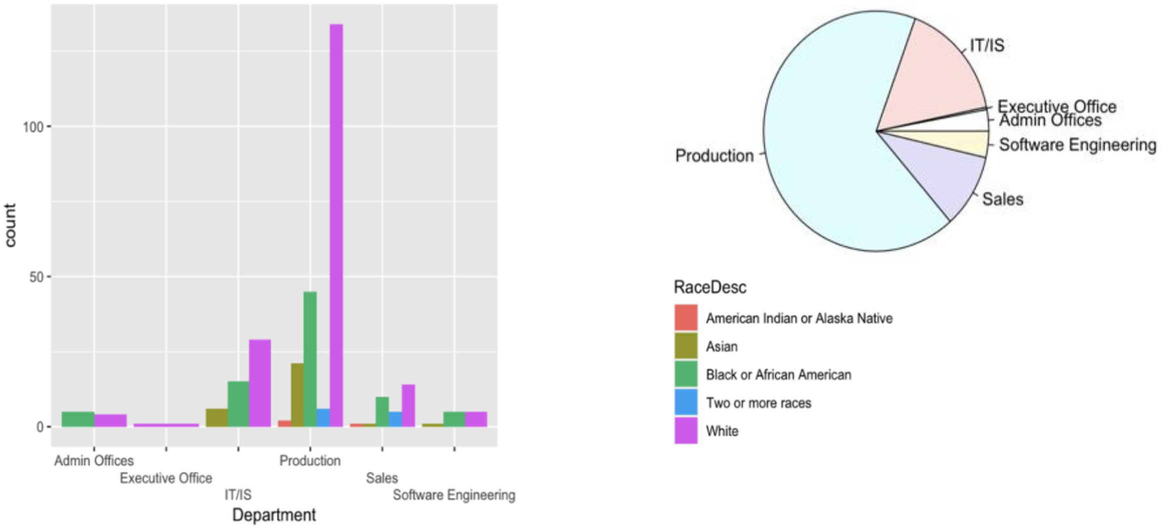


Figure 4: Racial Distribution per Department (left), No. Jobs per Department (top right)

Additional summarizing information about the data set can be found in Figure 13 to Figure 15 and Table 2 in the appendix.

2.2. Initial Regression Model

The standard approach is to start by fitting a linear regression model. It is one of the simplest and most powerful tools. Only if further analysis shows that such a model does not fit this use case more complex models shall be considered. The first step of the process of linear regression model fitting is an initial goodness of fit check alongside testing a few model assumptions. This is to ensure the chosen type of model is applicable to this data set before moving on to model selection methods. [6]

The goodness of fit test begins with checking for normality. Figure 5 shows the two plots used for this check. The histogram shows unimodularity and good symmetry. It shows heavy tails, but this is not necessarily problematic at this stage in the process. The qq-plot on the right shows overall good alignment with the 45 degrees line. Notably, even the ends do not trail off significantly. Instead, there is a horizontal arrangement of points between -0.25 and 0.25 of normal quantiles. To sum up, the plots do not show severe violations and the normality assumption is reasonable confirmed.

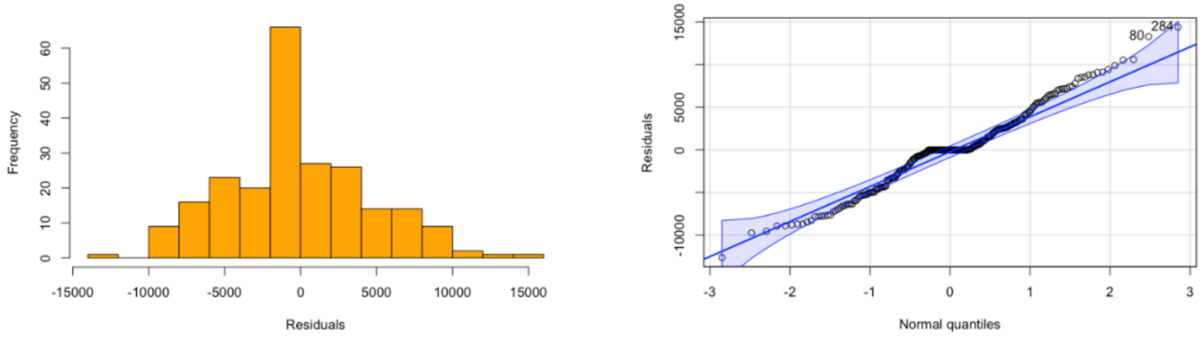


Figure 5: Histogram of Residuals (left) and qq-Plot (right), Full Model

Next, the constant variance assumption is checked. Figure 6 plots the residuals vs. fitted values. The points shall be randomly distributed around the zero-line and no pattern or trend should be visible [6]. The values very close to zero for fitted values above 100,000 can be explained with most of these data point to being unique for some predictors. Therefore, the model fits these points very specifically. This does not necessarily contradict the constant variance assumption. Together with the check for normality, we can conclude that no treatment of the response variable is necessary. This conclusion was further strengthened by implementing several transformation techniques which did not show improvement.

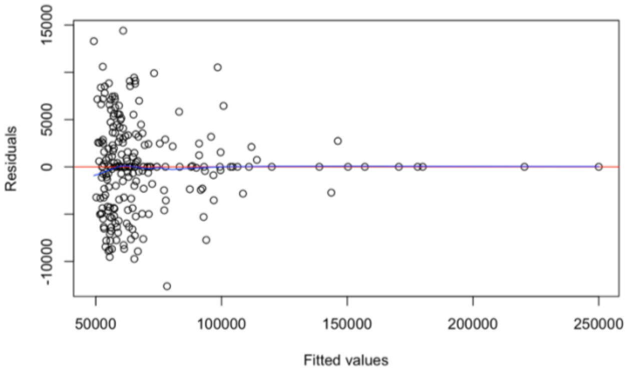


Figure 6: Residuals vs. Fitted Values, Full Model

To conclude this initial check of model assumptions. Figure 7 and Figure 8 show the residual plots for quantitative variables contained in the full model. For quantitative predictors, the linearity assumption shall hold. The check items are similar to those of constant variance. All variables are discrete but some with only very few possible values e.g. a satisfaction survey offering values between one and five. This makes the plots slightly harder to interpret. However overall, there are no clear indications of non-linearity or violations. This does not necessarily meet our expectations. As performance ratings or satisfaction statements are highly subjective, non-linearity in their relationship to salary was expected. See residual plots for all quantitative predictors in Figure 25 in the appendix.

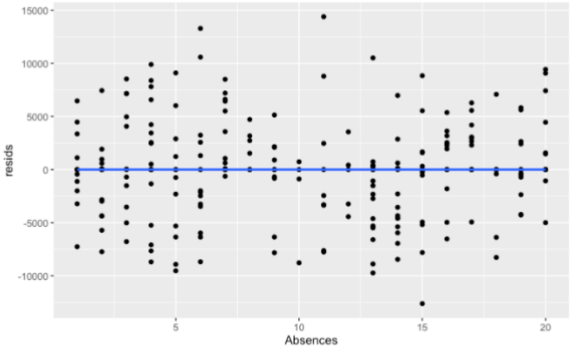


Figure 7: Residuals vs. Absences

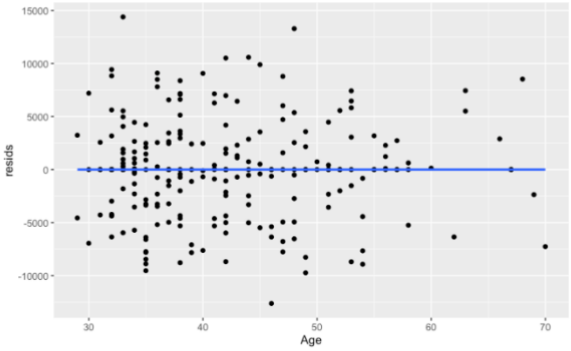


Figure 8: Residuals vs. Age

Outliers or multicollinearity are not checked prior to model selection. These topics will be discussed using the final model. To sum up, the initial goodness of fit test and model assumption check conclude positively and allow the analysis to move forward with model selection.

2.3. Model Selection

The full model has some obvious multicollinearity as some predictors are depicting the same information but on different levels of granularity. For example, each “Manager” should correlate with a “Department” which in turn contains certain “Positions”. But also, beyond issues of multicollinearity there exists the need for a formal model selection process. High dimensionality reduces the interpretation capabilities of individual model parameters. Furthermore, overfitted models show extremely high variance. This leads us to the bias vs. variance trade-off which is at the core of model selection. Large models tend to have low bias but high variance while reducing their size leads to lower variance, it increases bias. This can be expressed with the prediction risk. There are different approaches to estimating the prediction risk as a sum of the training risk and a complexity penalty. Furthermore, there are different algorithmic selection processes based on those estimators. This section will explore the most used methods.

The most holistic approach is called exhaustive search. Intuitively, this method computes all possible models and calculates at prediction risk estimation for each. However, 21 predictors allow for 2,097,152 theoretical models. This is a classic example of the exhaustive search being infeasible.

Before starting model selection, controlling and explanatory variables need to be defined. Controlling variables are declared as necessary and will be excluded from variable selection. This can be done to control sample bias or is sometimes necessary to support the research hypothesis. For this project, no variable is selected as controlling variables. The analysis can be viewed as a blank sheet approach maximizing the degrees of freedom during model selection.

2.3.1. Stepwise Regression

Stepwise regression is a greedy algorithm which can be used for model selection. It relies on the AIC. However, it is not guaranteed that it finds the model associated with the lowest AIC value. There are different types of stepwise regression based on adding or dropping variables. Forwards regression starts with a minimum model and adds variables. While backwards regression deselects variable from an initial full model. Finally, both directions can be combined as hybrid approach. Given the large number of predictors, it is often beneficial to use forward regression. However, for this project all three approaches are executed to gain the most insight of this technique. [6]

For forwards stepwise regression the minimum model including only the intercept shows a very high AIC just below 4700. Adding "Position" as a predictor reduces this value drastically to 4025. Subsequent models add four more predictors which only slightly reduces the AIC further to 4018. Figure 26 in the appendix shows the path of stepwise regression using the forward approach.

Furthermore, Figure 9 shows the path of stepwise regression using the backward approach. The full model shows an AIC of 4065. The algorithm continues to deselect as much as thirteen predictors. The first predictor to be removed is "State". This confirms impressions from the explanatory data analysis showing that there are no major differences between states. At the same time, "States" contains a large number of categories adding immense variance to the model. The absolute and relative decrease of AIC gets smaller with each step. Overall, it shows a more gradual evolution towards the final choice than forwards regression. The final model has an AIC of 4018.

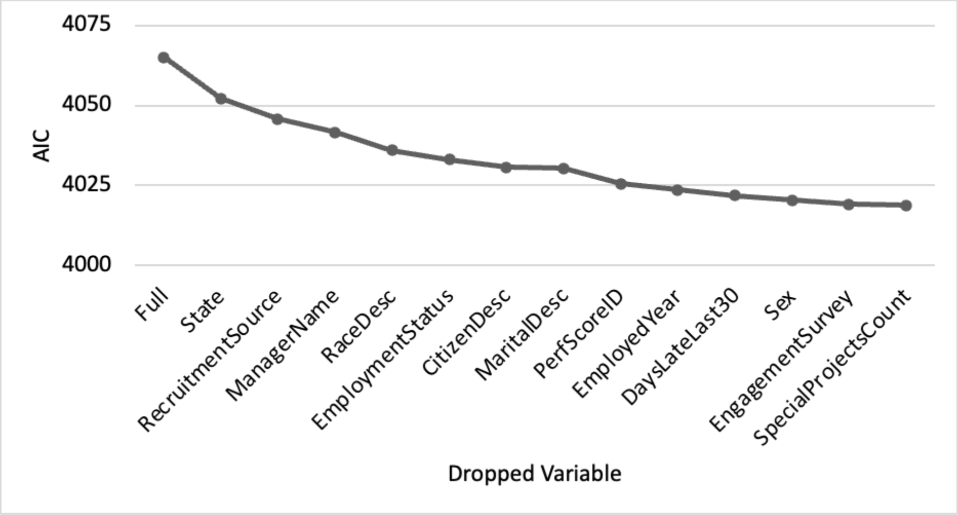


Figure 9: AIC Evolution through Stepwise Regression (Backward)

Finally, both way stepwise regression takes a very similar path compared to the backward approach. Indeed, it selects the same model as its preferred model (See Figure 27 in the appendix for the path of stepwise regression using the omnidirectional approach).

2.3.2. Regularized Regression

The group of regularized regression algorithm combine model fitting and selection algorithms [6]. In this project, Ridge, Lasso and Elastic Net regressions were implemented. As expected, Ridge regression did not down select any predictors. Lasso and Elastic Net regression showed similar but not identical results. Overall, both approaches selected more variables than stepwise regression. For all selections made during regularized regression, linear regression models were re-trained on the training data set. See Figure 28 in the appendix for associated plots.

2.3.3. Comparison

After applying different types of model selection methods, this section will evaluate and conclude by selecting a final model. Table 1 depicts all variables of the full model and shows which variables are selected by each model selection method. It becomes apparent how drastically the number of variables included in the model is reduced by every method. From initially 21 variables, stepwise and regularized regression reduce the number of variables till five to nine.

Table 1: Model Selection Overview

Variable	**Model**				
	Full, Ridge	Stepwise forward, both	Stepwise backward	Lasso regression	Elastic net
Marriedid	X	X			
Maritialdesc	X				
Sex	X				
Employmentstatus	X				
Department	X				
Perfscoreid	X			X	X
Recruitmentsource	X				
Salary	X				
Position	X	X	X	X	X
State	X			X	X
Age	X		X	X	X
Citizendesc	X				
Racedesc	X				
Hispaniclatino	X	X	X		
Employedyear	X			X	X
Managername	X			X	
Engagementsurvey	X				
Empsatisfaction	X	X	X	X	X
Specialprojectscount	X			X	
Dayslatelast30	X				
Absences	X	X	X	X	X

Each model selection method is differentiated by either the selection criteria used or the approach it uses. Therefore, it becomes necessary to compare selected models based on all available criteria used. Figure 10 depicts the different models, their number of predictors and adjusted R^2 values. Additionally, the three most commonly used selection criteria: Mallow's complexity penalty (Mallow's CP), Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). Even though the number of variables was reduced by more than 50%, the number of predictors is not proportionally reduced. This is because key variables with a large number of categories remain in the models. The adjusted R^2 values are all very similar slightly above 95%. However, models selected with stepwise regression show the highest values around 95.3% (See all adjusted R^2 values in Figure 29 in the appendix). Additionally, they achieve the lowest values for all three variable selection criteria. Based on these values, forward stepwise regression achieves slightly better values compared to backward and both way approaches.

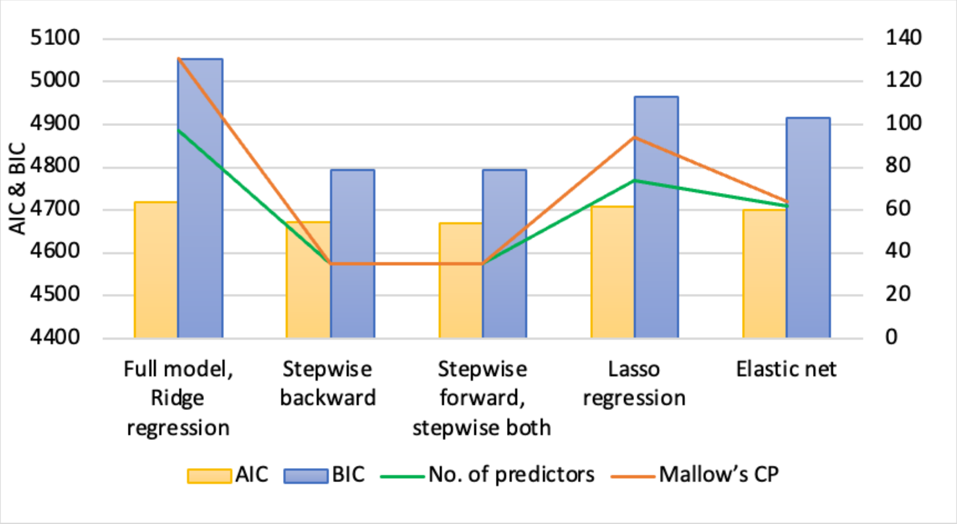


Figure 10: Variable Selection Criteria[1]

The second important group of indicators which need to be looked at during model selection are model performance indicators. Figure 11 depicts each model with values for the mean squared prediction error (MSPE), mean absolute prediction error (MAE), mean absolute percentage error (MAPE) and precision measure (PM). Backward stepwise regression achieves the lowest MSPE. However, all other performance indicators are preferable for regularized regression models with elastic net selecting the best model in each of the other three categories.

[1] See values in Table 3 in the appendix.

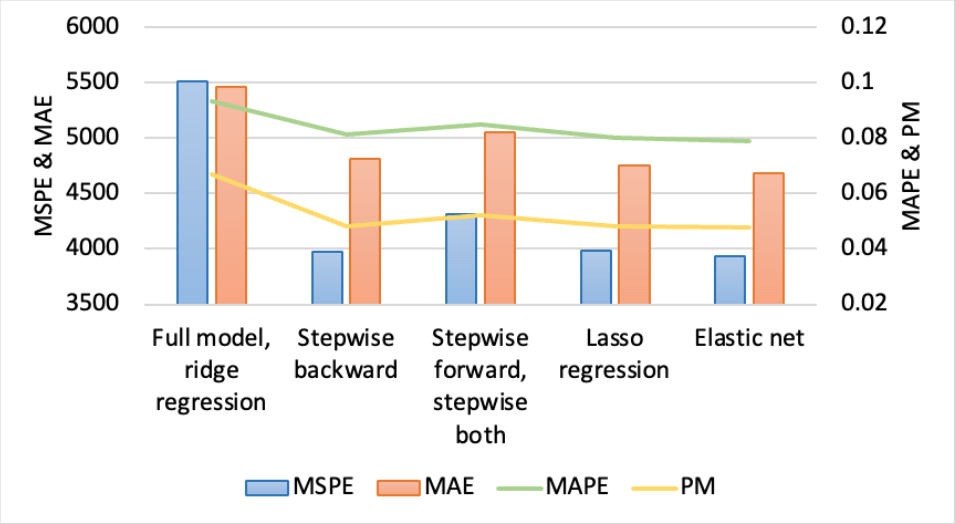


Figure 11: Model Performance Comparison[2]

2.4. Final Regression Model

As the conclusion in 2.3.3, we select the elastic net model as our final model based on its best overall performance indicators. It includes seven predictors, none of which is based on gender, race, or other personal backgrounds (See Table 5 in the appendix for the model summary). Additionally, it contains *State* which has the potential as a controlling variable as overall salary levels might differ from state to state.

In the following sections we go through checking goodness of fit and model assumptions again to ensure the final model can accurately reflect the data and result in accurate predictions. Then we check if multicollinearity issues happen in the final model and do experiments on removing outliers.

2.4.1. Goodness of Fit and Model Assumptions

The goodness of fit and model assumptions are assessed similarly to Section 2.2[3]. In Figure 30, the histogram of residuals for the elastic net model also demonstrates unipolarity and good symmetry as the full model does. The qq-plot shows overall good alignment with the 45-degree line and better result that the end of the line falls in the area at the 95% significance level compared to full models. Both plots show that the final model reasonably follows the normality assumption.

Regarding constant variance assumptions and uncorrelated errors, we observe a very similar picture as the full model. In Figure 31, the residuals randomly distributed around the zero-line and no clear pattern or trend, which confirms that both assumptions hold true.

As to be expected, Figure 32 shows that residuals of elastic net model vs. quantitative predictors plots shows the same random scatter around zero line as for the full model indicating there is no violation of linearity assumption.

Overall, the result is similar to the analysis of the full model. In conclusion, we can state that the normality, constant variance, uncorrelated errors, and linearity assumptions still hold true in the elastic net model

[2] See values in Table 4 in the appendix.

[3] Due to the similarity to Section 2.2, see Figure 30 to Figure 32 in the appendix.

based on the residual analysis. This proposes that the final model fits the human resource dataset and can reflect the data accurately.

2.4.2. Multicollinearity

Before testing the multicollinearity, we have checked for the correlation coefficient which is useful to evaluate the linear relationship between the response variable and of the predicting variables and helps in detecting the multicollinearity. The maximum correlation value among the numeric variables comes out to be 0.298 that is between the *EmpSatisfaction* and *Employee PerfScoreID.* See the entire correlation matrix in Table 6 in the appendix. Overall, the values are insignificant and do not indicate any strong linear relationship between individual predictors.

When computing the variance inflation factor (VIF) we found an error of having aliased coefficient in the Net Elastic Model, which means some variables are perfectly correlated. It turned out that one data point was singular war both *State* and *Position* and had caused this issue. To proceed, we remove this entry from the dataset. The VIF Threshold value came out to be 27.54 for the final model. All VIF values adjusted for their degrees of freedom are close to one and therefore lower than the threshold (See all VIF values in Table 7 in the appendix). We conclude that multicollinearity is not apparent between selected predictors.

2.4.3. Outliers

Figure 12 depicts the Cook's Distance plot which can be used to identify potential outliers. There are several approaches for defining a threshold. The red line shows the threshold for distance values D_i depending on the number of observations n of $D_i > \frac{4}{n}$. This threshold of 0.017 would identify eight potential outliers. However, given the general roughness of the plot only observation 201 with a Cook's distance of around 0.08 was further investigated. No model characteristic showed a significant change upon its removal. Thus, we concluded that it is neither an outlier, nor influential and kept the observation in the training data set.

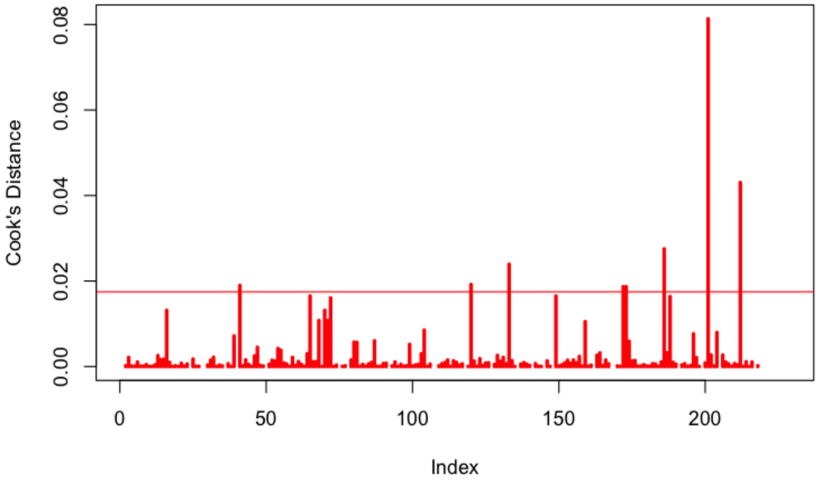


Figure 12: Cook's Distance Plot

3. Results

To sum up, the linear full model including all predictors almost satisfies the normality and constant variance assumption. Through the process of model selection, we conclude that forward stepwise regression gives the highest R-Square as well as the lowest Mallow's Cp, AIC and BIC, meaning that it fits the dataset well and has less information loss compared to other models. Elastic net has the lowest MSPE, MAE and MAPE meaning it has higher prediction accuracy than other models. We select the Elastic net model due to its superior overall performance. The normality and residual plots of Elastic net model also show improvements compared to the initial full model. Further investigation showed that there neither significant multicollinearity nor outliers. The coefficient of determination (R^2) is 0.9609. Thus, the model explains 96.1% of the variability in the salary of the employees.

The predictors included in the finals model are: *State*, *Position*, *EmployedYear*, *EmpSatisfaction*, *Absences*, *PerfScoreID* and *Age*. The following statements are made holding all other predictors constant. The two states with the highest coefficients are PA and CA. However, due to the low number of observations in those states, this shall not be interpreted as causality. Other predictors show more intuitive coefficients. With each year spent with the company the salary increases by $92. With each level of higher employer satisfaction their salary increases by $497. Less intuitive are the coefficients of positive $136 for absence days and -$32 for higher performance scores. This clearly shows the limitation of a multiple linear regression model in its ability of explaining causal relationships.

Since none of the models selected *Sex* as a predictor it is reasonable to conclude that there might not be gender biases in salary. This is to be deemed very positive insight and contrary to initially stated concerns about today's salary structures. This fact complements insights about a strong female representation in higher management roles within the company.

Similarly, *Race* was not selected during any model selection process. However, this cannot lead to a positive conclusion as *Hispaniclatino* was selected by both forward and backward stepwise regression. In the model selected by forward stepwise regression Hispanics and Latinos are earning $3,107 less holding all other predictors constant. Though the interpretation for the selection is limited, this could hint at racial biases. Against this concern one might argue that Hispanics and Latinos only account for a small fraction of the overall employee population (10%). This potentially could have caused such results. We propose that the company continue to pay attention on gender and racial balance on future hires. For all these reasons, it is beneficial that the final regression model avoided such predicators.

Furthermore, other than essential predictors like *Position* and *State*, the final model does not include predictors containing categories with very few observations as discussed during exploratory data analysis in Section 2.1. Still, the categorical variable *Position* contains 31 categories and has strong predictive power but might cause model overfitting. A possible point for improvement might be to derive more aggregated position level such as junior and senior employee, manager and director.

4. Discussion and Outlook

One challenge of this project was a clear separation between focusing on predictive and explanatory purposes. It is unlikely, that a company would rely solely on a regression model based on their current salary data to define salaries for future hires. This might only be applicable to fast-growing companies which have yet to define a well-documented process for the definition of salaries. In their case, such an analysis might yield a first starting point in the definition process. Even then, it remains questionable whether a companies would opt to implement a regression model which does not offer clear marginal relationships.

Instead, it is likely that human resource business partners desire a salary structure which has easy to understand relationships. For a large cooperation with an established salary definition process, the purpose of such analysis might be more driven by challenging their current salary structure. To assess whether salaries are actually determined without biases in terms of gender or other personal backgrounds.

If the model was to be used for assessing salaries of new hires, some information would be missing. At the point of hiring there would be no history of performance assessments or the employee’s punctuality. Therefore, baseline values for new hires need to be defined which result in neutral assessment prior to actual data collection within the company.

Further research could expand the analysis across many different companies to assess wider trends. A comparison among different companies can also increase the understanding of data within one specific company. And most importantly, more evenly distributed data among underrepresented categories could significantly improve model performance.

Bibliography

[1] "Data-driven Automated Decision-Making in Assessing Employee Performance and Productivity: Designing and Implementing Workforce Metrics and Analytics," *Psychosociological Issues in Human Resource Management,* vol. 07, no. 2, pp. 13-18, 2019.

[2] . H. W. Lee and M. C. Lin, "A study of salary satisfaction and job enthusiasm–mediating effects of psychological contract," *Applied Financial Economics,* vol. 24, pp. 1577-1583, 2014.

[3] L. Zhang, " Gender and racial gaps in earnings among recent college graduates," *The Review of Higher Education,* vol. 32, pp. 51-72, 2008.

[4] J. W. Curtis, "Faculty salary equity: still a gender gap?," *On Campus with Women,* vol. 39, 2010.

[5] "Human Resources Data Set, Dataset used for learning data visualization and basic regression," Kaggle.com, 29 04 2021. [Online]. Available: https://www.kaggle.com/rhuebner/human-resources-data-set. [Accessed 05 10 2021].

[6] D. G. Kleinbaum and L. L. Kupper, Applied regression analysis and other multivariable methods, North Scituate: Duxbury Press, 2013.

Appendix

Figures

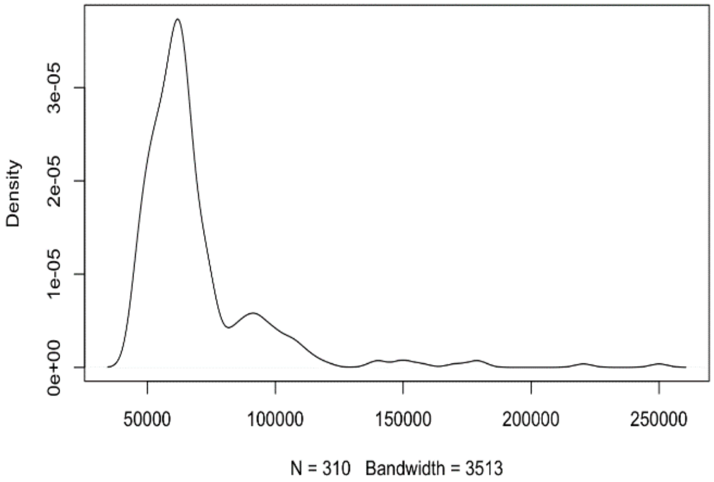


Figure 13: Salary Density

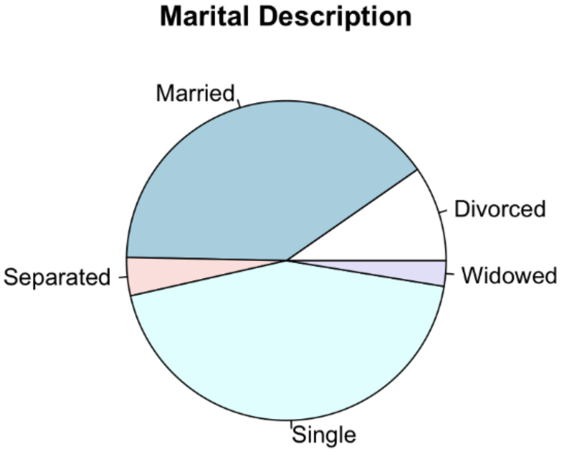


Figure 14: Pie Chart, Marital Description

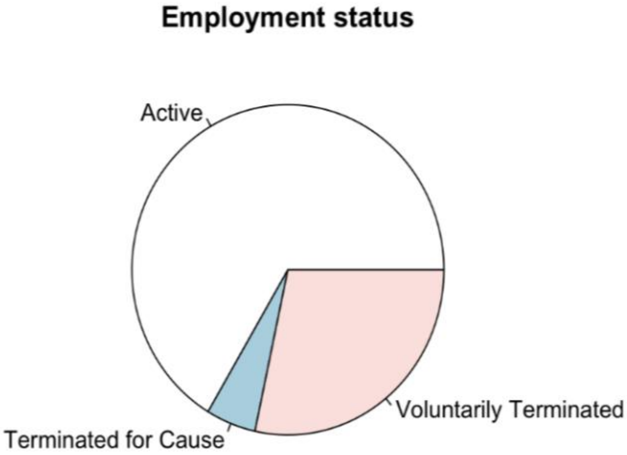


Figure 15: Pie Chart, Employment Status

Box-Plots

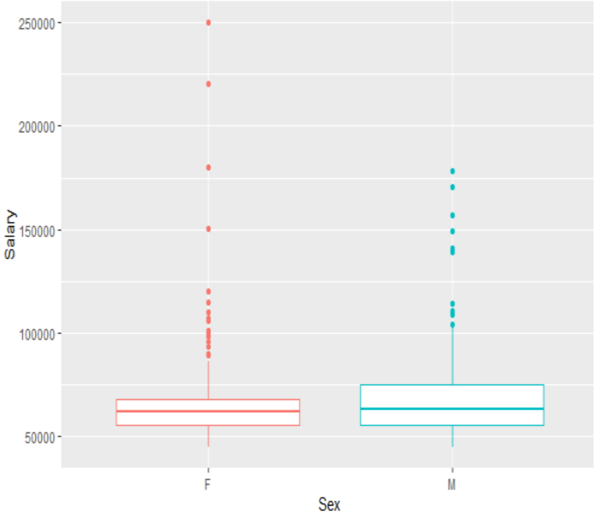


Figure 16: Boxplot, Salary vs. Sex

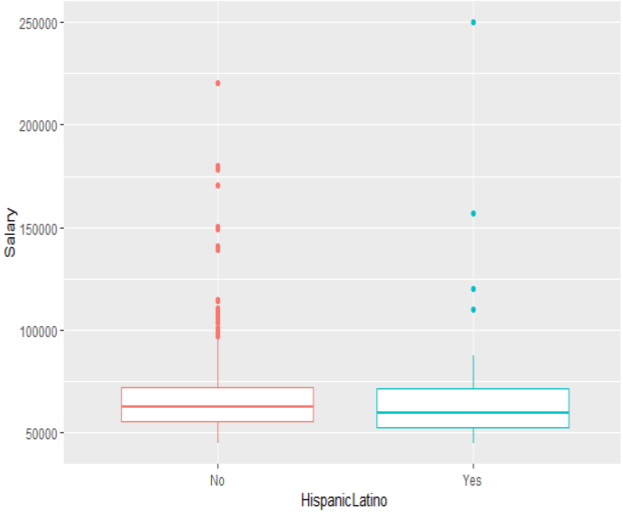


Figure 17: Boxplot, Salary vs. HispanicLatino

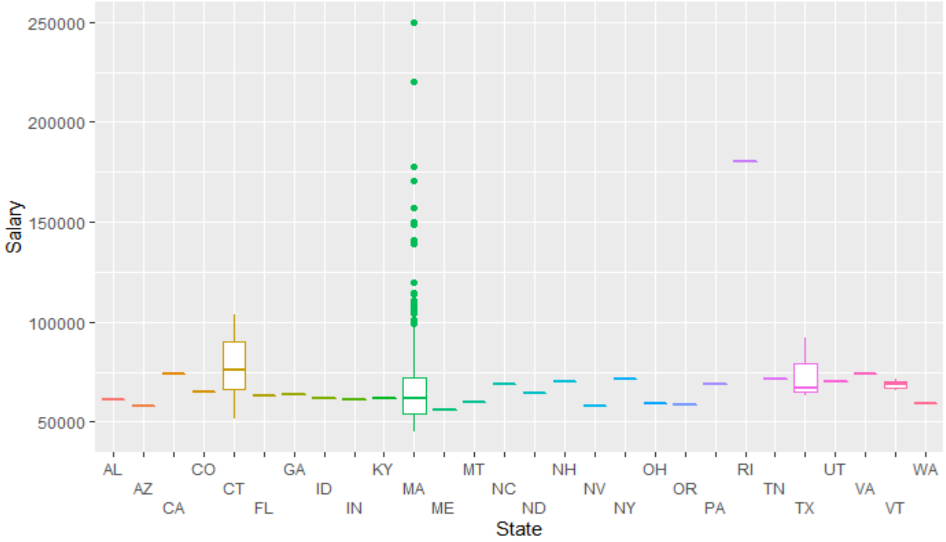


Figure 18: Boxplot, Salary vs. State

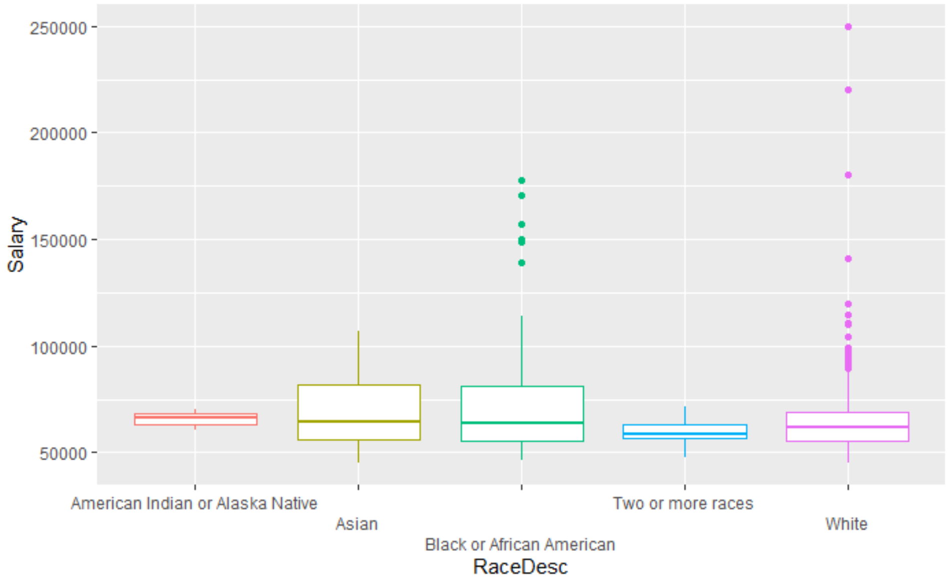


Figure 19: Boxplot, Salary vs. RaceDesc

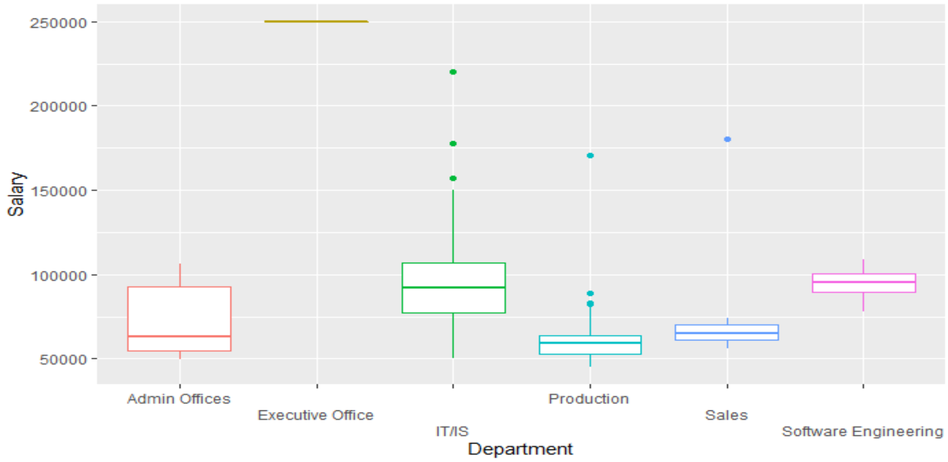


Figure 20: Boxplot, Salary vs. Department

Scatter Plots

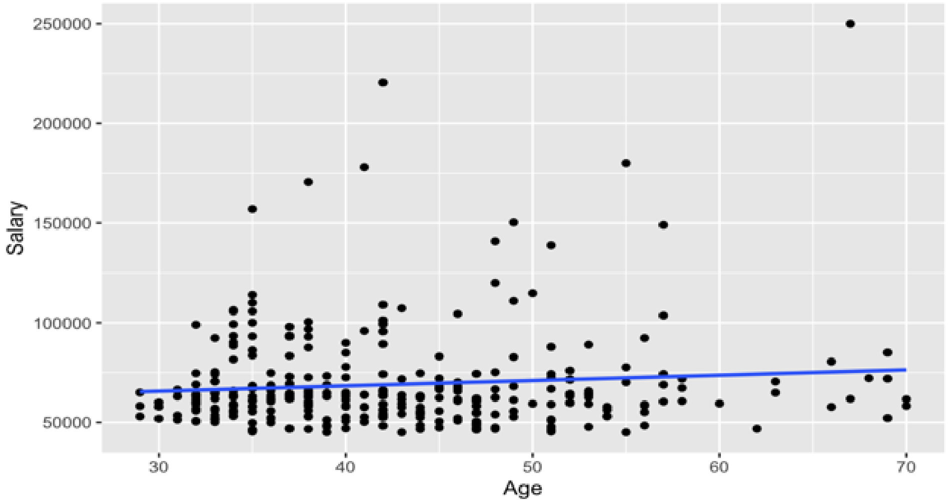


Figure 21: Scatter Plot, Salary vs. Age

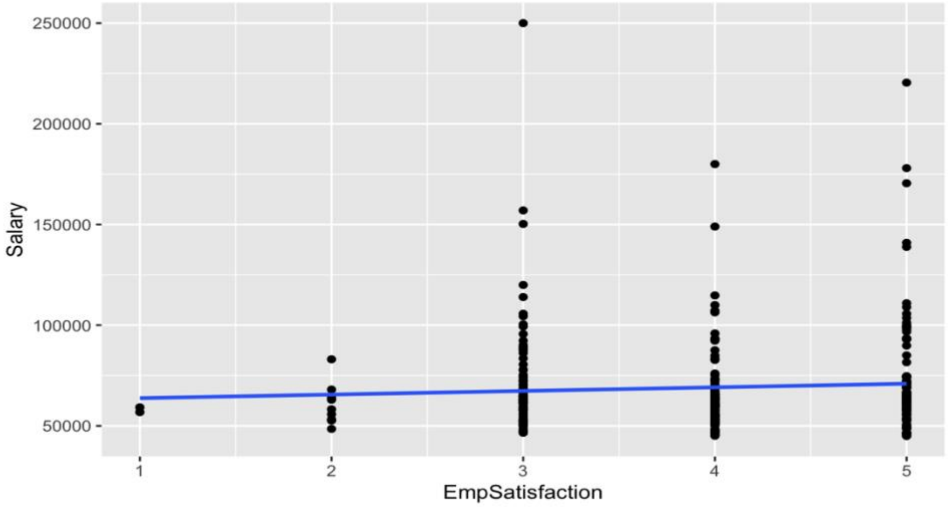


Figure 22: Scatter Plot, Salary vs. EmpSatisfaction

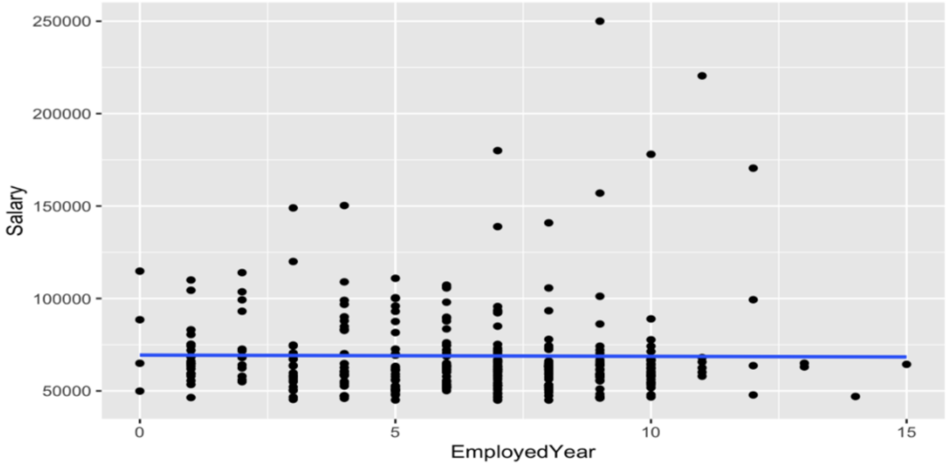


Figure 23: Scatter Plot, Salary vs. EmployedYear

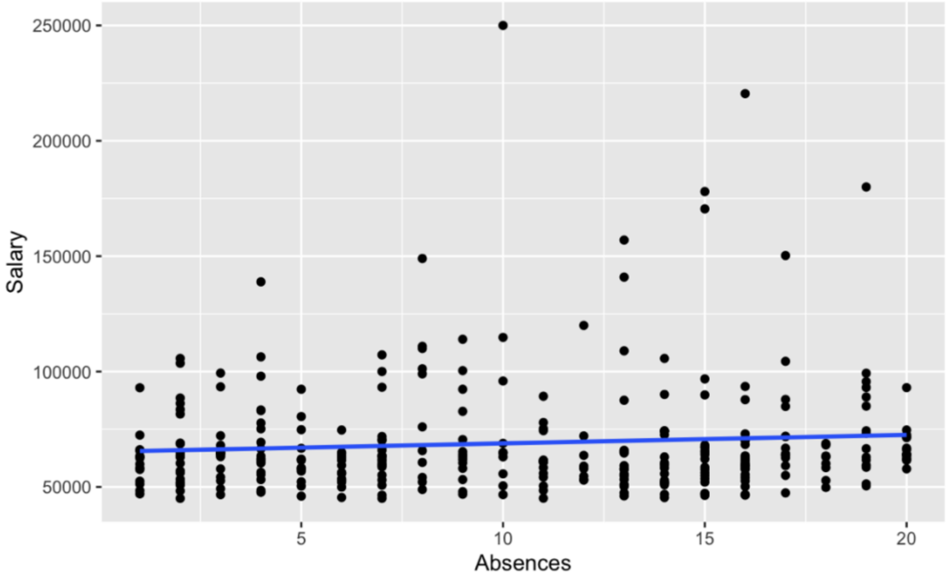


Figure 24: Scatter Plot, Salary vs. Absences

Initial Linear Regression Model

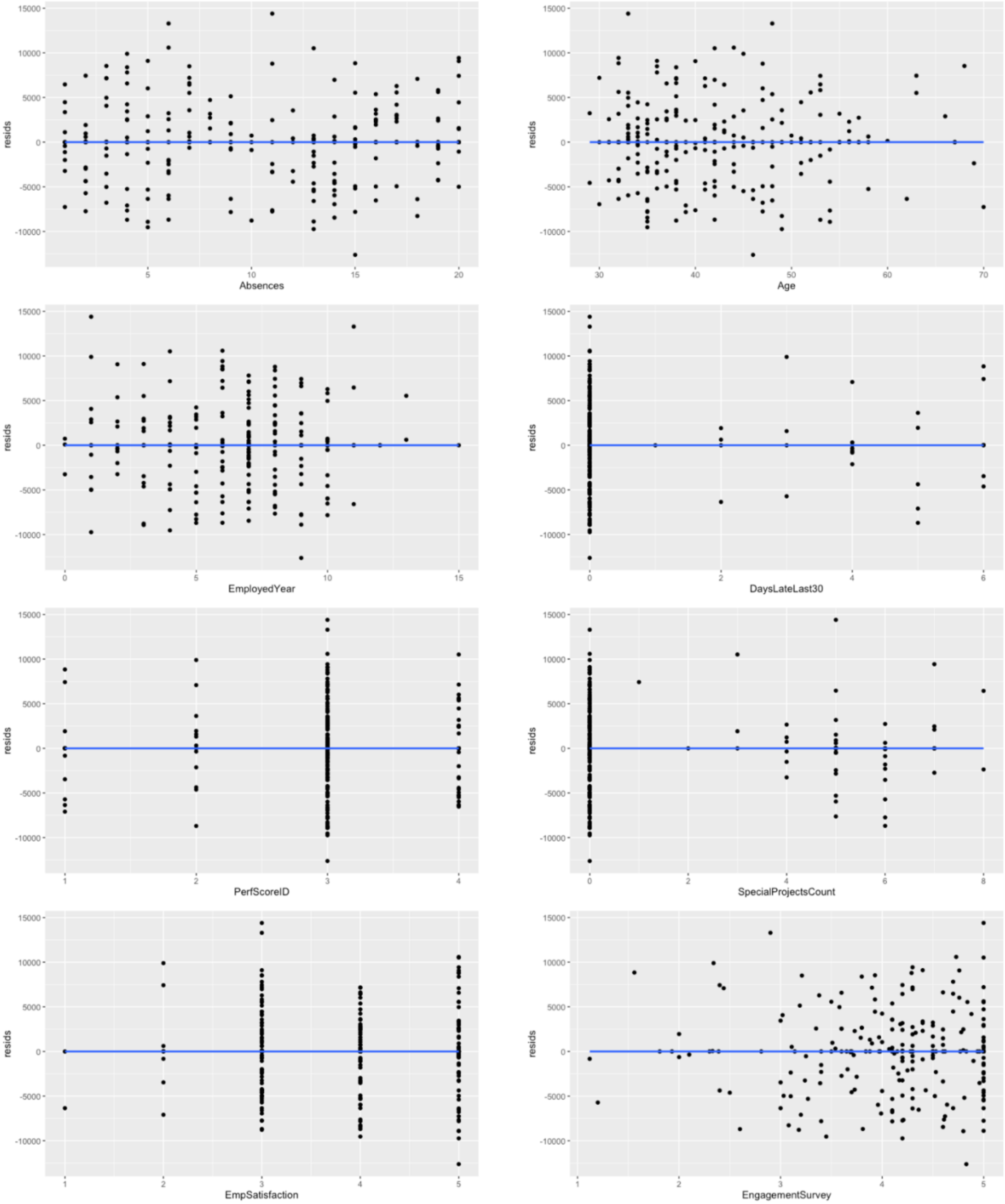


Figure 25: Residual Analysis for quantitative Predictors, Full Model

Model Selection

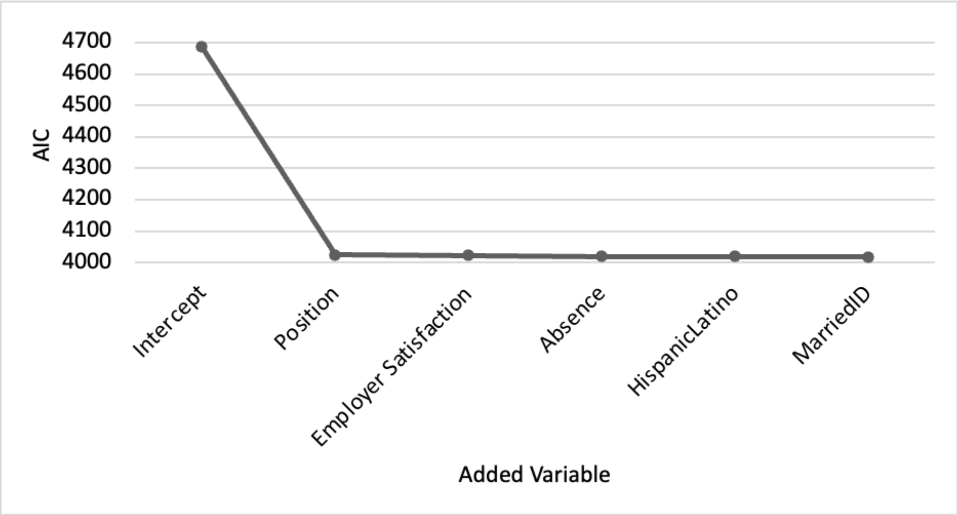


Figure 26: AIC Evolution through Stepwise Regression (Forward)

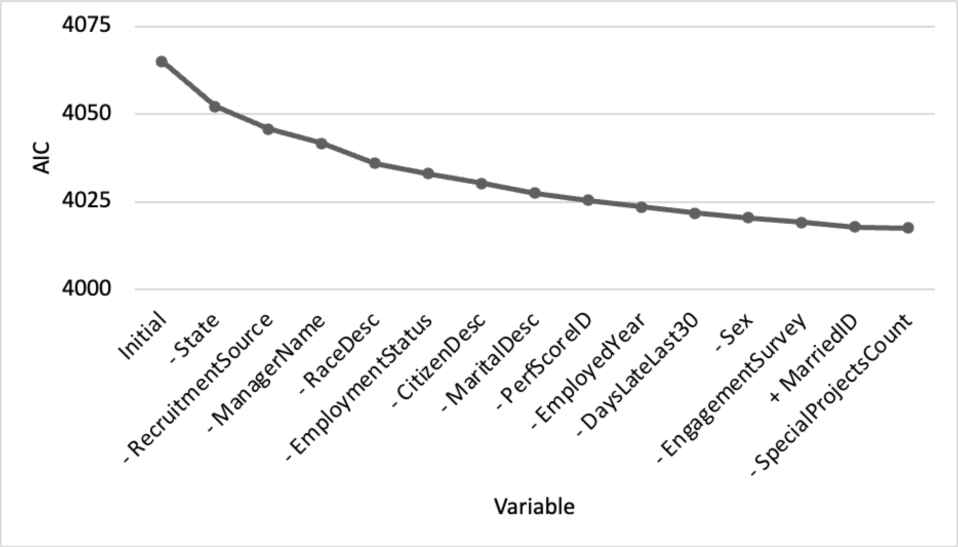


Figure 27: AIC Evolution through Stepwise Regression (Both)

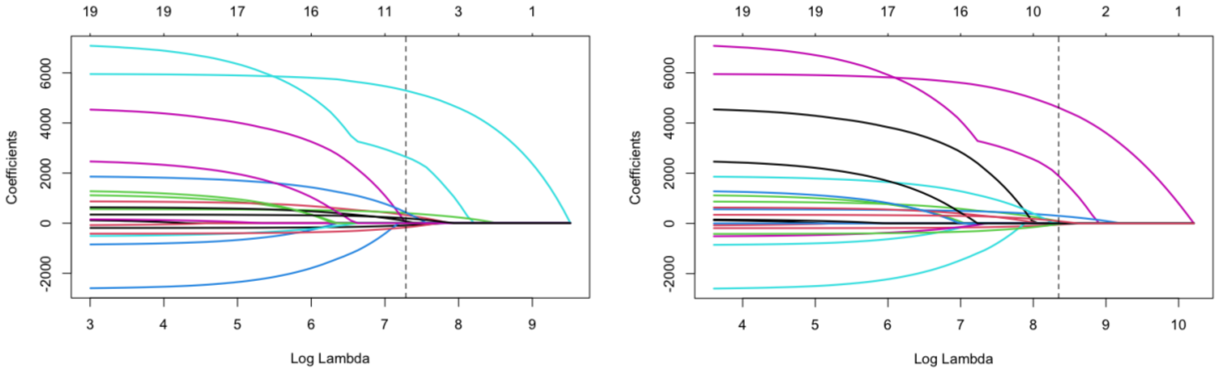


Figure 28: Regularized Regression, Lasso (left) and Elastic Net (right)

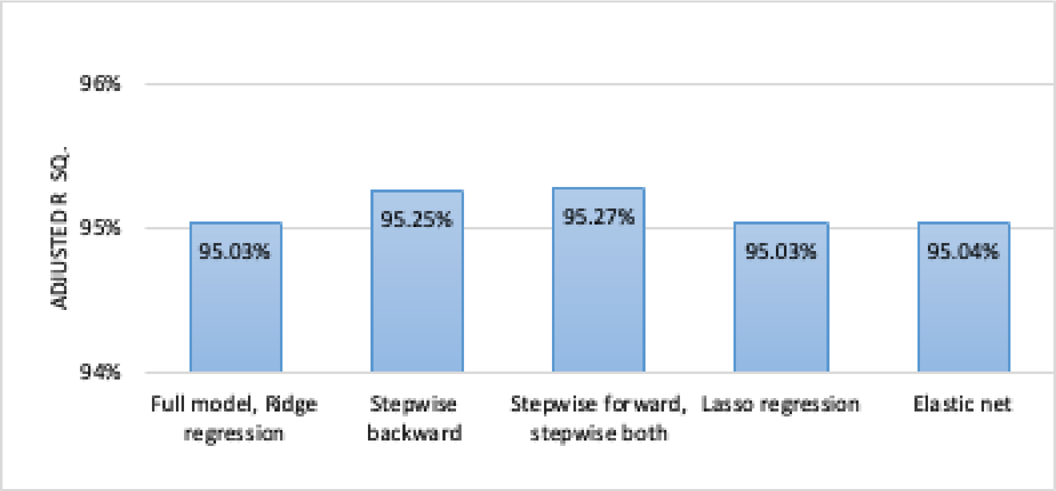


Figure 29: Adjusted R-Squared Values, Model Comparison

Final Linear Regression Model

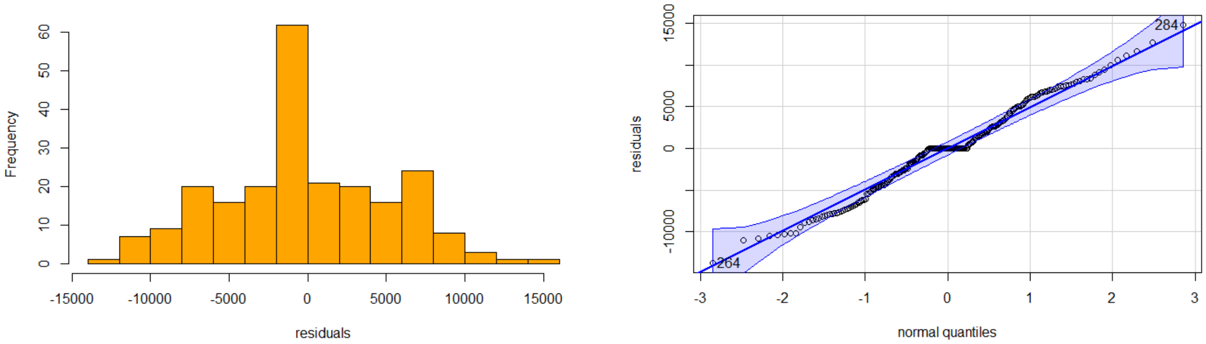


Figure 30: Histogram of Residuals (left) and qq-Plot (right), Elastic Net Model

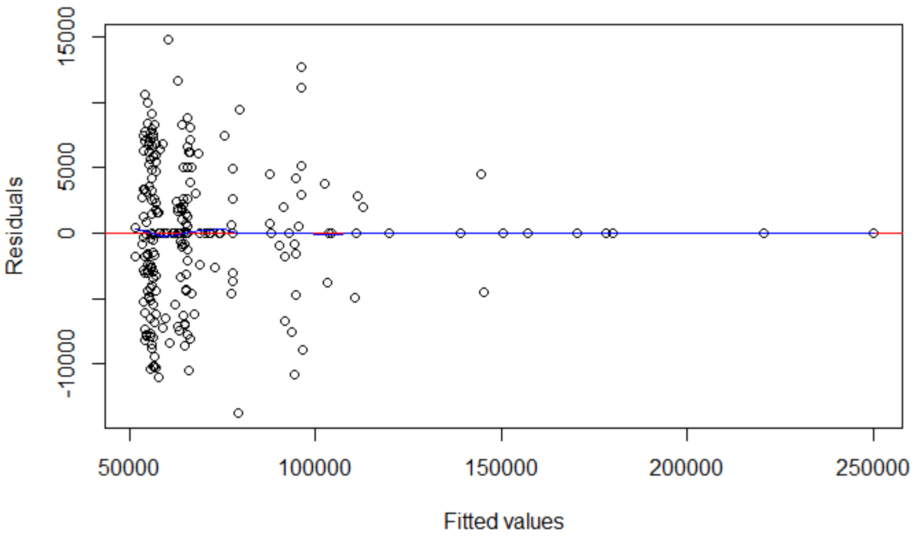


Figure 31: Residuals vs. Fitted Values Plots, Elastic Net Model

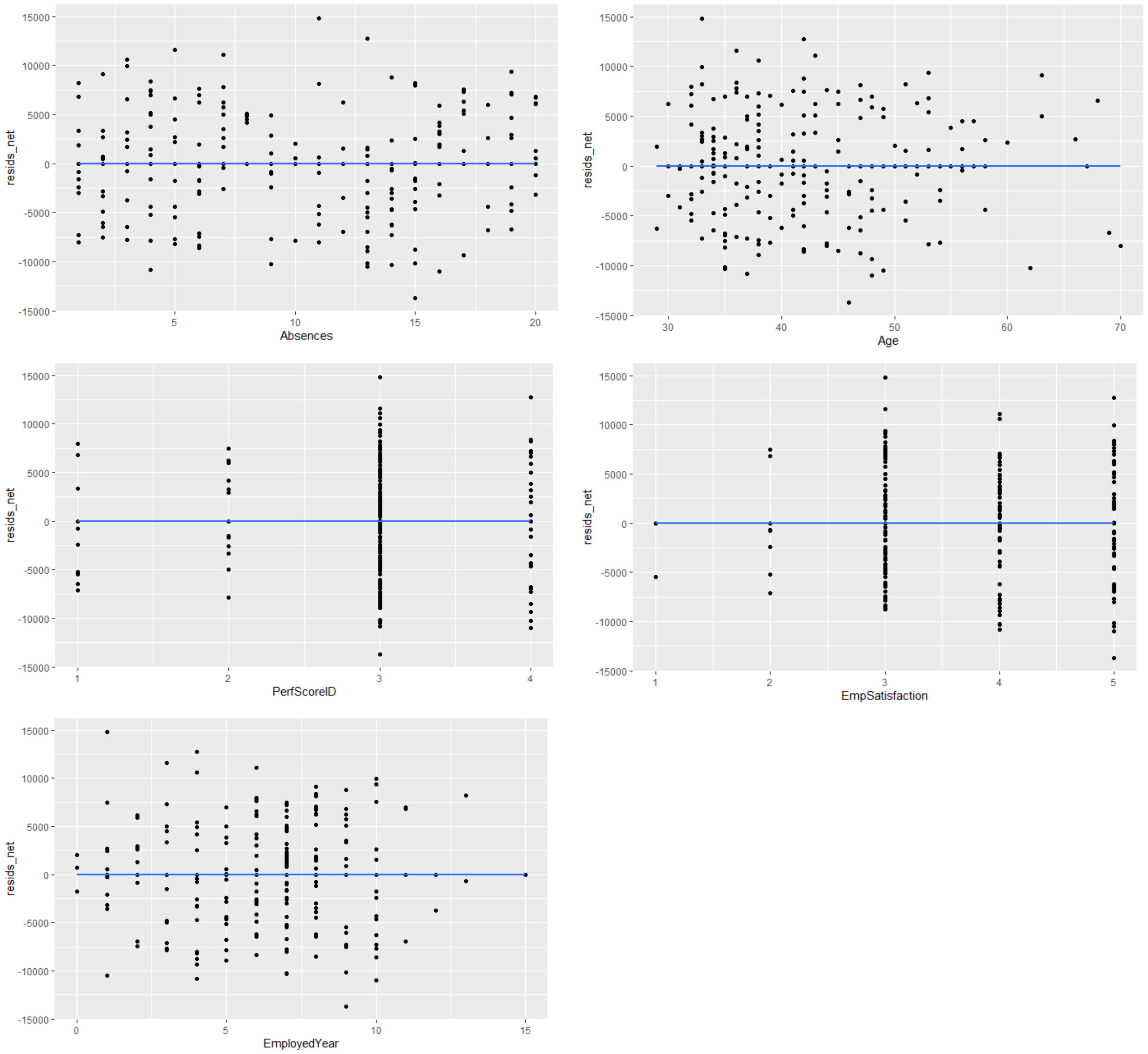
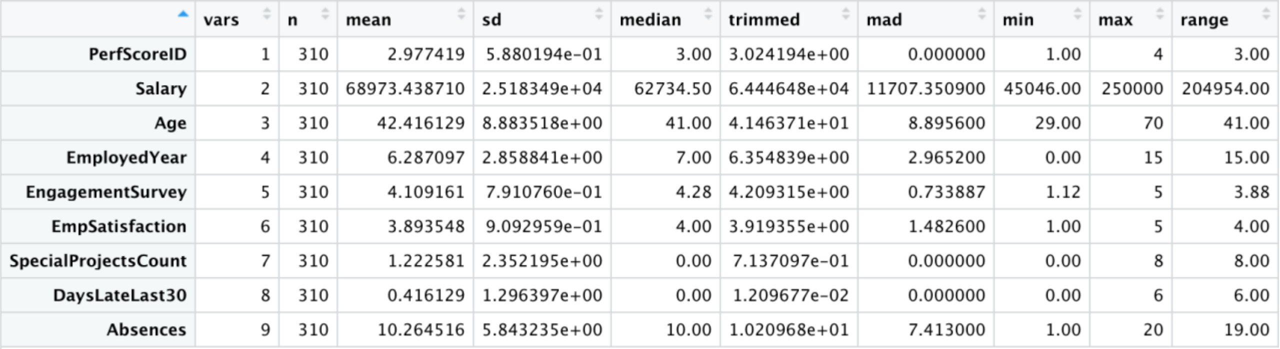


Figure 32: Residual Analysis for Quantitative Predictors, Elastic Net Model

Tables

Table 2: Numerical Analysis of HR Data Set



	vars	n	mean	sd	median	trimmed	mad	min	max	range
PerfScoreID	1	310	2.977419	5.880194e-01	3.00	3.024194e+00	0.000000	1.00	4	3.00
Salary	2	310	68973.438710	2.518349e+04	62734.50	6.444648e+04	11707.350900	45046.00	250000	204954.00
Age	3	310	42.416129	8.883518e+00	41.00	4.146371e+01	8.895600	29.00	70	41.00
EmployedYear	4	310	6.287097	2.858841e+00	7.00	6.354839e+00	2.965200	0.00	15	15.00
EngagementSurvey	5	310	4.109161	7.910760e-01	4.28	4.209315e+00	0.733887	1.12	5	3.88
EmpSatisfaction	6	310	3.893548	9.092959e-01	4.00	3.919355e+00	1.482600	1.00	5	4.00
SpecialProjectsCount	7	310	1.222581	2.352195e+00	0.00	7.137097e-01	0.000000	0.00	8	8.00
DaysLateLast30	8	310	0.416129	1.296397e+00	0.00	1.209677e-02	0.000000	0.00	6	6.00
Absences	9	310	10.264516	5.843235e+00	10.00	1.020968e+01	7.413000	1.00	20	19.00

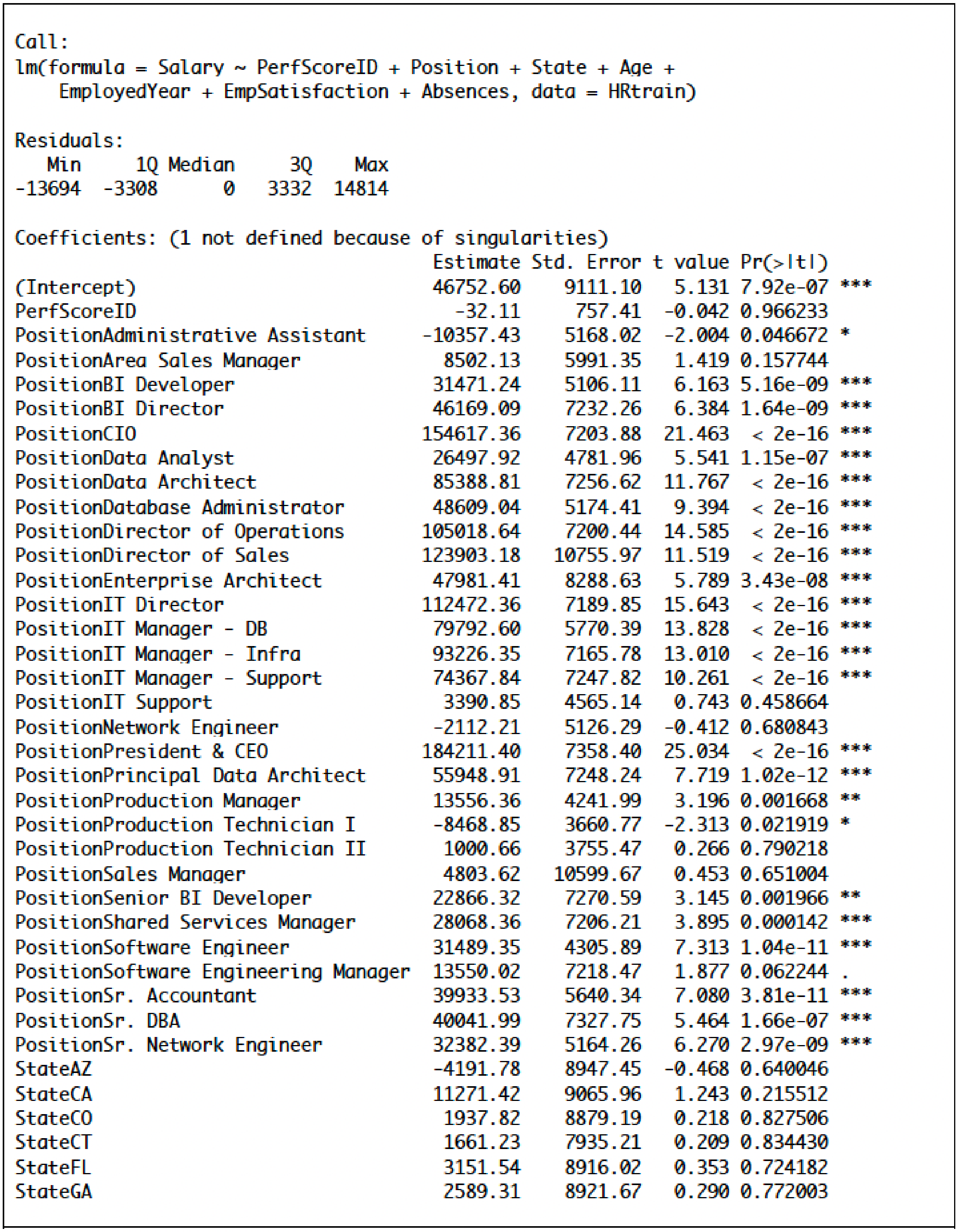
Table 3: Variable Selection Criteria

Regression model	**Num. of predictors**	**Adjusted R^2**	**Mallow's CP**	**AIC**	**BIC**
Full model, Ridge regression	97	0.9503	131	4717	5053
Stepwise backward	35	0.9525	26	4671	4794
Stepwise forward, stepwise both	35	0.9527	25	4670	4793
Lasso regression	74	0.9503	94	4707	4965
Elastic net	62	0.9504	64	4700	4916

Table 4: Model Performance Comparison

Regression model	**MSPE**	**MAE**	**MAPE**	**PM**
Full model, ridge regression	55078189	5456	0.0931	0.0668
Stepwise backward	39726001	4817	0.0814	0.0482
Stepwise forward, stepwise both	43125080	5048	0.0849	0.0523
Lasso regression	39829696	4756	0.0802	0.0483
Elastic net	39309436	4687	0.0787	0.0477

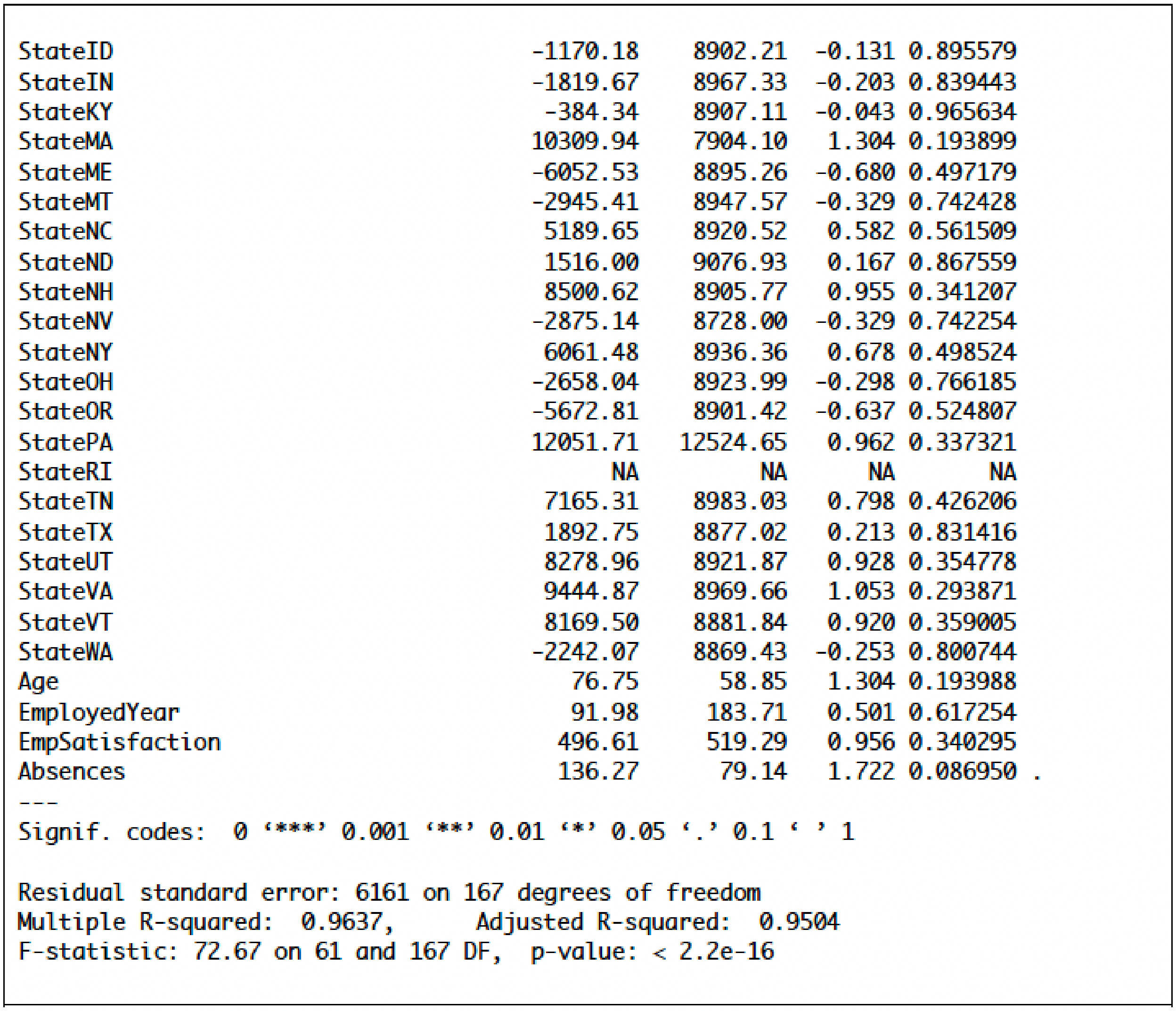
Table 5: Final Model Summary



```
Call:
lm(formula = Salary ~ PerfScoreID + Position + State + Age + 
    EmployedYear + EmpSatisfaction + Absences, data = HRtrain)

Residuals:
   Min     1Q Median     3Q    Max 
-13694  -3308      0   3332  14814 

Coefficients: (1 not defined because of singularities)
                                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)                           46752.60    9111.10   5.131 7.92e-07 ***
PerfScoreID                             -32.11     757.41  -0.042 0.966233    
PositionAdministrative Assistant     -10357.43    5168.02  -2.004 0.046672 *  
PositionArea Sales Manager             8502.13    5991.35   1.419 0.157744    
PositionBI Developer                  31471.24    5106.11   6.163 5.16e-09 ***
PositionBI Director                   46169.09    7232.26   6.384 1.64e-09 ***
PositionCIO                          154617.36    7203.88  21.463  < 2e-16 ***
PositionData Analyst                  26497.92    4781.96   5.541 1.15e-07 ***
PositionData Architect                85388.81    7256.62  11.767  < 2e-16 ***
PositionDatabase Administrator        48609.04    5174.41   9.394  < 2e-16 ***
PositionDirector of Operations       105018.64    7200.44  14.585  < 2e-16 ***
PositionDirector of Sales            123903.18   10755.97  11.519  < 2e-16 ***
PositionEnterprise Architect          47981.41    8288.63   5.789 3.43e-08 ***
PositionIT Director                  112472.36    7189.85  15.643  < 2e-16 ***
PositionIT Manager - DB               79792.60    5770.39  13.828  < 2e-16 ***
PositionIT Manager - Infra            93226.35    7165.78  13.010  < 2e-16 ***
PositionIT Manager - Support          74367.84    7247.82  10.261  < 2e-16 ***
PositionIT Support                     3390.85    4565.14   0.743 0.458664    
PositionNetwork Engineer              -2112.21    5126.29  -0.412 0.680843    
PositionPresident & CEO              184211.40    7358.40  25.034  < 2e-16 ***
PositionPrincipal Data Architect      55948.91    7248.24   7.719 1.02e-12 ***
PositionProduction Manager            13556.36    4241.99   3.196 0.001668 ** 
PositionProduction Technician I       -8468.85    3660.77  -2.313 0.021919 *  
PositionProduction Technician II       1000.66    3755.47   0.266 0.790218    
PositionSales Manager                  4803.62   10599.67   0.453 0.651004    
PositionSenior BI Developer           22866.32    7270.59   3.145 0.001966 ** 
PositionShared Services Manager       28068.36    7206.21   3.895 0.000142 ***
PositionSoftware Engineer             31489.35    4305.89   7.313 1.04e-11 ***
PositionSoftware Engineering Manager  13550.02    7218.47   1.877 0.062244 .  
PositionSr. Accountant                39933.53    5640.34   7.080 3.81e-11 ***
PositionSr. DBA                       40041.99    7327.75   5.464 1.66e-07 ***
PositionSr. Network Engineer          32382.39    5164.26   6.270 2.97e-09 ***
StateAZ                               -4191.78    8947.45  -0.468 0.640046    
StateCA                               11271.42    9065.96   1.243 0.215512    
StateCO                                1937.82    8879.19   0.218 0.827506    
StateCT                                1661.23    7935.21   0.209 0.834430    
StateFL                                3151.54    8916.02   0.353 0.724182    
StateGA                                2589.31    8921.67   0.290 0.772003    
```



```
StateID                              -1170.18    8902.21  -0.131 0.895579
StateIN                              -1819.67    8967.33  -0.203 0.839443
StateKY                               -384.34    8907.11  -0.043 0.965634
StateMA                              10309.94    7904.10   1.304 0.193899
StateME                              -6052.53    8895.26  -0.680 0.497179
StateMT                              -2945.41    8947.57  -0.329 0.742428
StateNC                               5189.65    8920.52   0.582 0.561509
StateND                               1516.00    9076.93   0.167 0.867559
StateNH                               8500.62    8905.77   0.955 0.341207
StateNV                              -2875.14    8728.00  -0.329 0.742254
StateNY                               6061.48    8936.36   0.678 0.498524
StateOH                              -2658.04    8923.99  -0.298 0.766185
StateOR                              -5672.81    8901.42  -0.637 0.524807
StatePA                              12051.71   12524.65   0.962 0.337321
StateRI                                    NA         NA      NA       NA
StateTN                               7165.31    8983.03   0.798 0.426206
StateTX                               1892.75    8877.02   0.213 0.831416
StateUT                               8278.96    8921.87   0.928 0.354778
StateVA                               9444.87    8969.66   1.053 0.293871
StateVT                               8169.50    8881.84   0.920 0.359005
StateWA                              -2242.07    8869.43  -0.253 0.800744
Age                                     76.75      58.85   1.304 0.193988
EmployedYear                            91.98     183.71   0.501 0.617254
EmpSatisfaction                        496.61     519.29   0.956 0.340295
Absences                               136.27      79.14   1.722 0.086950 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6161 on 167 degrees of freedom
Multiple R-squared:  0.9637,	Adjusted R-squared:  0.9504 
F-statistic: 72.67 on 61 and 167 DF,  p-value: < 2.2e-16
```

Table 6: Correlation Coefficient Matrix

Predictor	**Perfscore**	**Age**	**Employed year**	**Emp satisfaction**	**Absences**	**Salary**
Perfscoreid	1.000	0.099	0.115	0.298	0.080	0.146
Age	0.099	1.000	-0.021	-0.053	-0.037	0.175
Employedyear	0.115	-0.021	1.000	0.009	0.003	0.025
Empsatisfaction	0.298	-0.053	0.009	1.000	0.090	0.092
Absences	0.080	-0.037	0.003	0.090	1.000	0.093
Salary	0.146	0.175	0.025	0.092	0.093	1.000

Table 7: Variance Inflation Factor

Predictors	**Gvif**	**Df**	**Gvif^(1/(2*df))**
Perfscoreid	1.419725	1	1.191522
State	239.3705	26	1.111095
Position	235.6763	29	1.098758
Age	1.641773	1	1.281317
Employedyear	1.612118	1	1.269692
Empsatisfaction	1.456007	1	1.206651
Absences	1.331421	1	1.153872