# Final Project - Algorithms for Decision Making

Noah Faas, Given Sandamela, JonJeng Thao

11/26/2024

##Code

## Initial Report and Slides

Introduction:

Question: How can we predict the precipitation in a city on a given day based on various geographic and other weather-related factors like elevation, wind, and distance to a coast?

It is important to make predictions about precipitation to know how to prepare for upcoming weather. To learn more about the weather prediction process and make our own predictions, we will be researching how we can predict the precipitation in a city on a given day based on various geographic and other weather-related factors like elevation, wind, and distance to a coast.

Our dataset is sourced from the United States National Weather Service, and includes 16 months worth of both forecasted and observed weather data from 167 cities across the United States from January of 2021 to June of 2022. This data is freely provided by the weather.gov website for transparency and research like this. Each observation in the dataset represents a forecast from 12 hours before the actual temperature was observed. The variables that the original dataset contained were date, city, state, whether the forecast was predicting the high or low temperature, the amount of hours prior to the observed temperature that the forecast was made, the observed temperature, the forecasted temperature, the observed precipitation, forecasted characteristics of the weather, and whether the row may contain an error. The original dataset contains 651,968 rows.

To clean and wrangle our dataset, we first joined the weather dataset itself with another csv containing definitions for each of the forecast outlook acronyms in the main dataset to make the data easier to interpret. We also added a column representing the residual between observed temp and forecasted temp in case that could be an interesting explanatory variable. Lastly, we decided that we only wanted to look at forecasts from 12 hours before the observation and that we only wanted to see high temperature values instead of both high and low to keep the methodology of observations consistent. We chose to only use weather data from Texas for our modeling and analysis because our models seem to struggle with the full dataset. Finally, we filtered out all observations where there is a possible error in the data and dropped all missing values to maintain the highest possible accuracy of our model. We now have 6,054 rows left in our cleaned dataset.

talk about the filtering better.

We will be trying to create a model that will predict the observed precipitation `log_observed_precip` based on 10 explanatory variables. These variables are all shown below.

Table 1: Variables (1 of 3)

| Variable | log_observed_precip | observed_temp | forecast_temp |
|---|---|---|---|
| Description | Log of actual precipitation + 1 | Actual temperature | Predicted temperature |
| Type | Numerical | Numerical | Numerical |

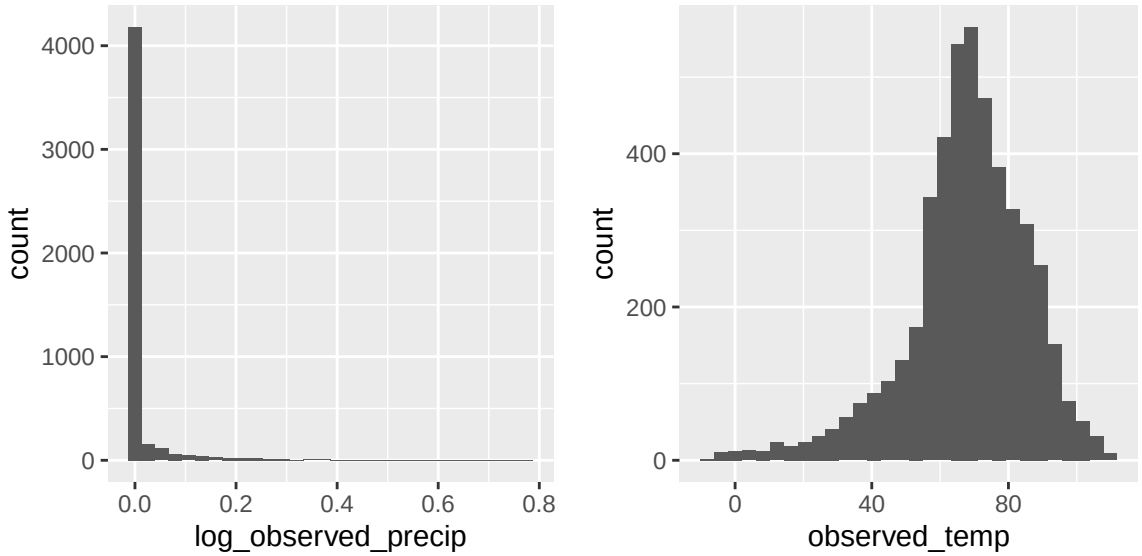
Table 2: Variables (2 of 3)

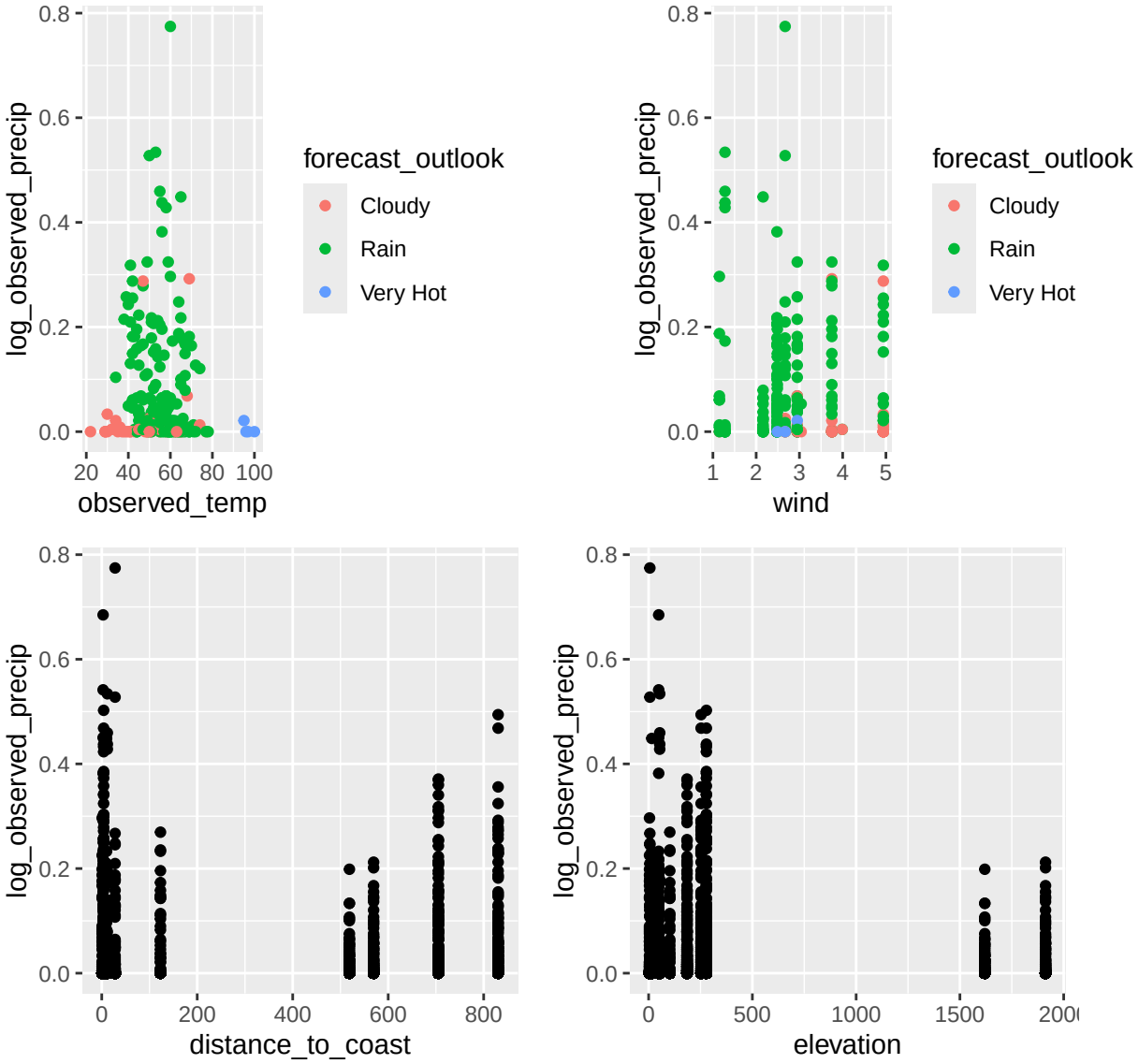
| Variable | lat | koppen | elevation | distance_to_coast |
|---|---|---|---|---|
| Description | Latitude of city | Köppen Climate Classification | Elevation of city | Distance from city to nearest coast |
| Type | Numerical | Categorical | Numerical | Numerical |

Table 3: Variables (3 of 3)

| Variable | wind | avg_annual_precip | forecast_residual | forecast_outlook |
|---|---|---|---|---|
| Description | Average annual wind speed of city | Average annual precipitation of city | Absolute value of the residual between predicted and actual temperature | Predicted weather characteristics |
| Type | Numerical | Numerical | Numerical | Categorical |

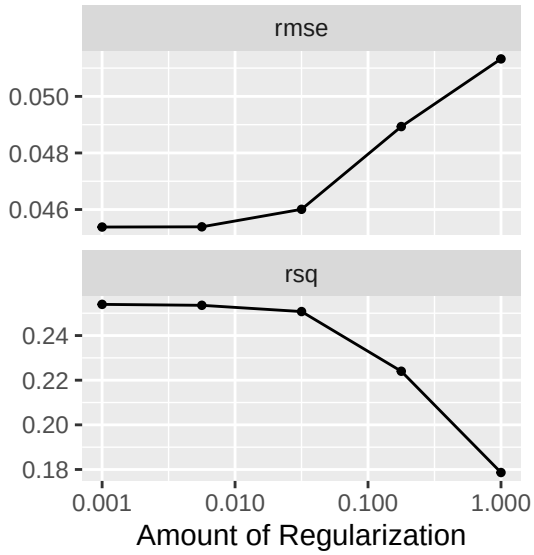
Exploratory Data Analysis:






Based on our exploratory data analysis, we found that observed precipitation is skewed right and observed temperature is skewed left. Due to the extreme right skew of our response variable, we have chosen to transform observed_precip into a new variable to model instead, log_observed_precip, which is calculated by log10(observed_precip + 1). In terms of relationships, we also found that observed precipitation seems to be normally distributed with regard to observed temperature at a center of about 60 degrees. Wind does not seem to be very highly correlated with precipitation, at least visually. Based on our scatterplot of precipitation and distance to coast, it seems that greater distance from a coast tends to be linked to decreased precipitation, which makes sense intuitively since rain comes from bodies of water. Lastly, similar to the previous relationship, it seems that elevation is also negatively correlated with precipitation.

Model building:



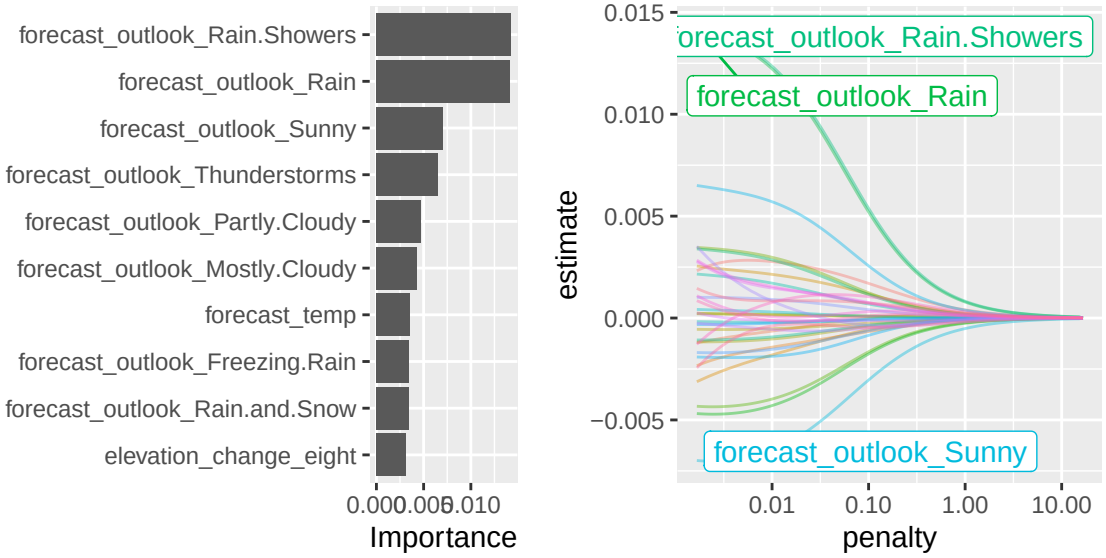

##Ridge Regression Model:

We used tune grid to optimize the values of penalty of the ridge model used to predict the log observed precipitation. Once we created the ridge model, we selected the best penalty which will maximize our r-squared value which we found to be 0.00001. We Then fit our ridge model on our training table, and use the testing table to calculate the optimal r-squared value which came up to 28.5%

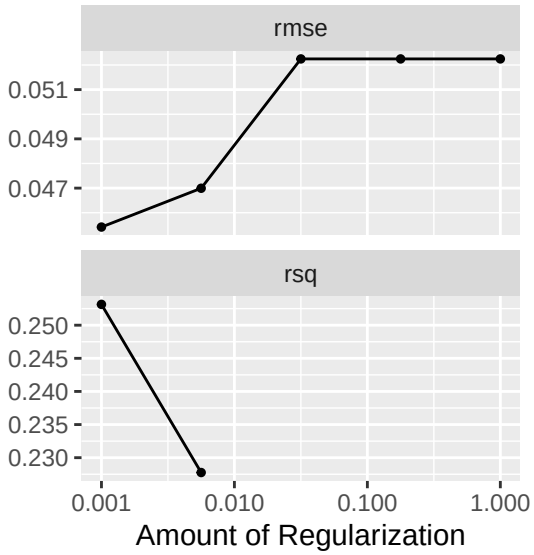
```
#> # A tibble: 1 x 2
#>   penalty .config
#>     <dbl> <chr>
#> 1   0.001 Preprocessor1_Model1
```

```
#> # A tibble: 1 x 3
#>   .metric .estimator .estimate
#>   <chr>   <chr>          <dbl>
#> 1 rsq     standard       0.285
```

```
#> # A tibble: 5 x 3
#>   term                           estimate penalty
#>   <chr>                             <dbl>   <dbl>
#> 1 forecast_outlook_Rain.Showers   0.0142    0.001
#> 2 forecast_outlook_Rain           0.0141    0.001
#> 3 forecast_outlook_Thunderstorms  0.00650   0.001
#> 4 forecast_temp                   0.00352   0.001
#> 5 forecast_outlook_Freezing.Rain  0.00346   0.001
```




This is a plot of the coefficients in the ridge model. As we increase the penalty, each coefficient gets closer to zero, but may not have a value of zero. In this model, we get the best penalty equal to 0.00001. Our final fit model has an r^2 equal to 28.5. We get that the 5 most important variables are 'forecast_outlook_Rain.Showers,forecast_outlook_Rain,forecast_outlook_Rain.Sunny,forecast_outlook_Thunderstorms,foreca Forecast outlook is a categorical. varible, and using our = ridge model we can see that it is the most significant variable.




##Lasso Regression Model:

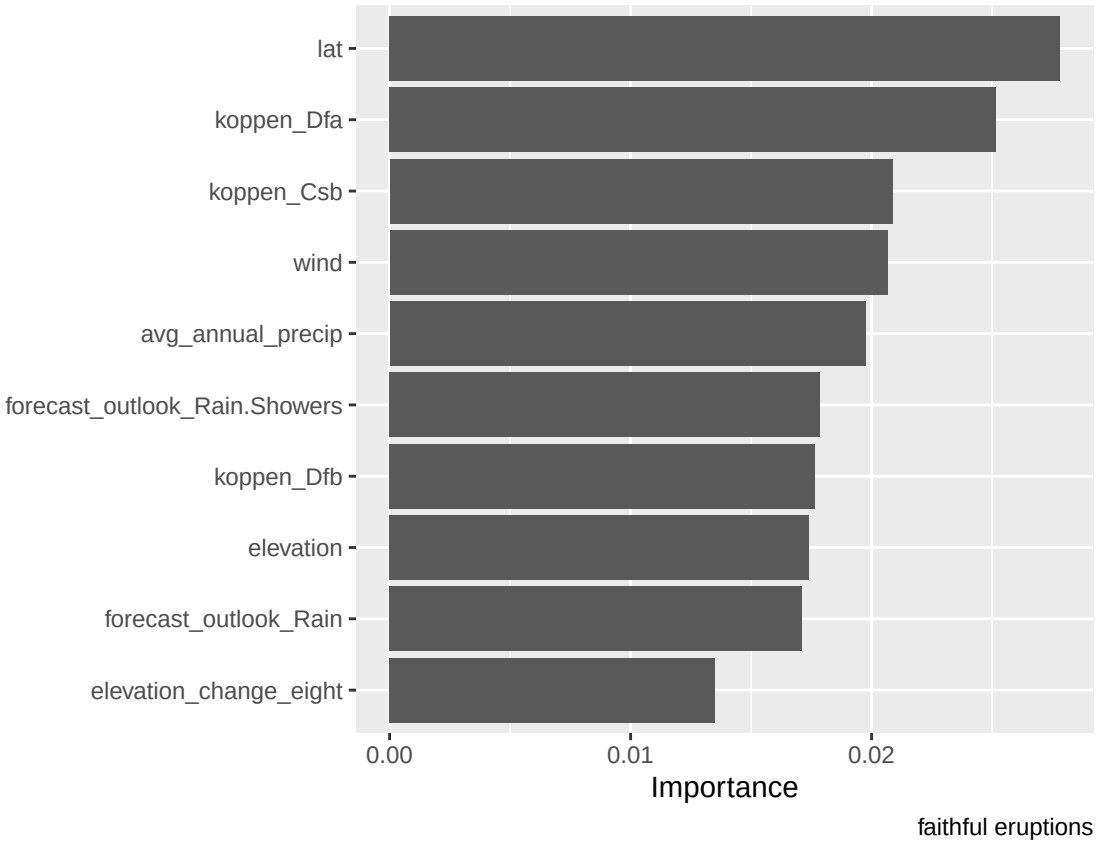
This is the tune results for the lasso model when trying out different values for the penalty. In this case, we want to maximize the r-squared.

```
#> # A tibble: 1 x 2
#>   penalty .config
#>     <dbl> <chr>
#> 1   0.001 Preprocessor1_Model1
```

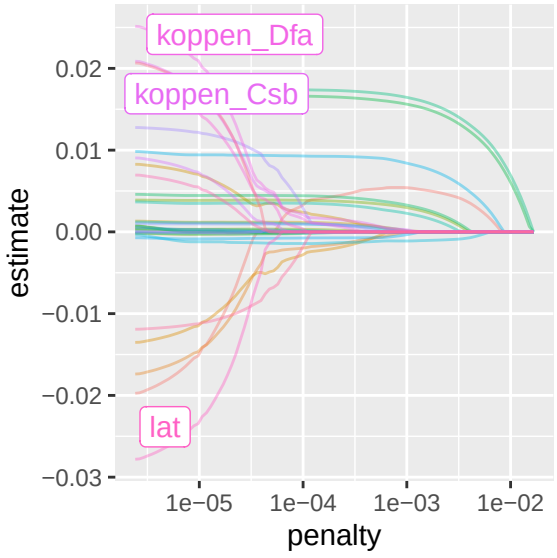
```
#> # A tibble: 1 x 3
#>   .metric .estimator .estimate
#>   <chr>   <chr>          <dbl>
```

```
#> 1 rsq     standard       0.283
```

```
#> # A tibble: 14 x 3
#>    term                              estimate penalty
#>    <chr>                                <dbl>   <dbl>
#>  1 forecast_outlook_Rain.Showers   0.0164      0.001
#>  2 forecast_outlook_Rain           0.0156      0.001
#>  3 forecast_outlook_Thunderstorms  0.00840     0.001
#>  4 avg_annual_precip               0.00541     0.001
#>  5 forecast_outlook_Rain.and.Snow  0.00328     0.001
#>  6 forecast_outlook_Freezing.Rain  0.00297     0.001
#>  7 forecast_outlook_Snow           0.00227     0.001
#>  8 forecast_outlook_Sunny         -0.00111     0.001
#>  9 forecast_outlook_Very.Cold     -0.000339    0.001
#> 10 koppen_Csb                      0.000257    0.001
#> 11 forecast_residual               0.000163    0.001
#> 12 forecast_outlook_Cloudy         0.000125    0.001
#> 13 elevation_change_four           0.000117    0.001
#> 14 elevation                      -0.0000137   0.001
```




Using the vip() function, we determined that `observed_temp` is the variable with the most importance in the model. This means that this coefficient has the most impact in predicting `log_observed_precip`.




This graph displays the lasso model for the data set. Notice that similarly to the ridge model, the lines also get closer to zero, but instead of rarely actually having a value of zero, the lasso model makes sure the coefficient is zero when there is zero variance. Our least comlex model has only the explanatory variable `forecast_ooutlook_Thunderstorms`

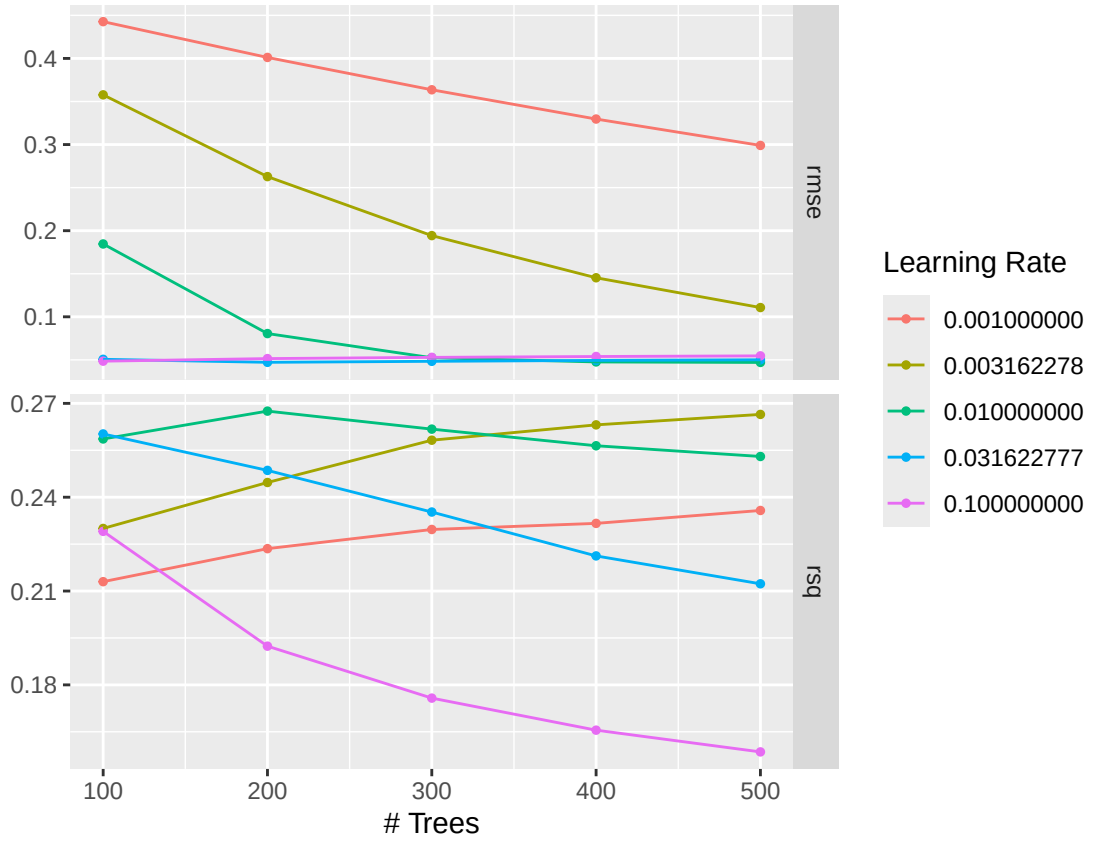
## Optimizing the cost complexity (finding Optimal Tree)

This plot shows the optimal value for cost complexity. In this case, we are interested in selecting the cost complexity with the least value. We will select by one standard error because it makes the model less complex.

We removed this and just give the explanation.

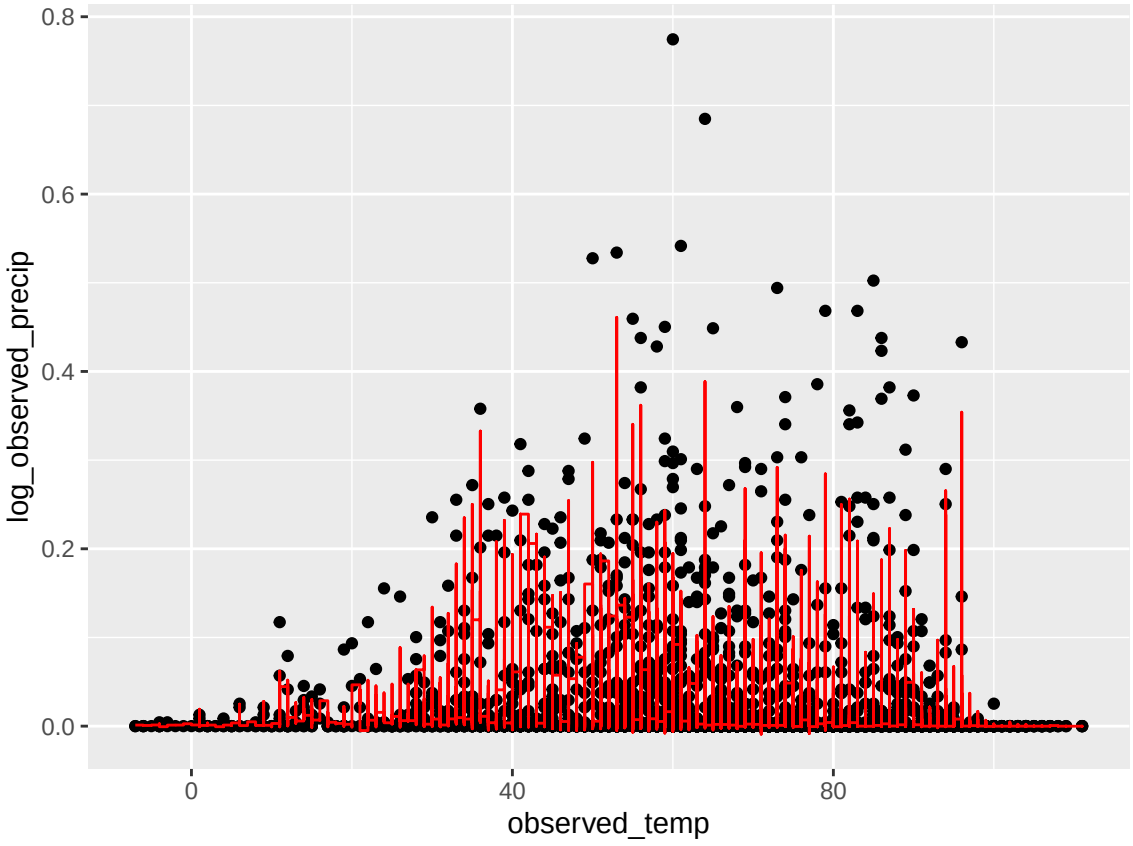
A cost complexity of 0.1 had the lowest rmse with a value of 0.0618 and rsq of 0.154.

#forest boosting




```
#> # A tibble: 1 x 3
#>   trees learn_rate .config
#>   <int>      <dbl> <chr>
#> 1   100        0.1 Preprocessor1_Model21
```

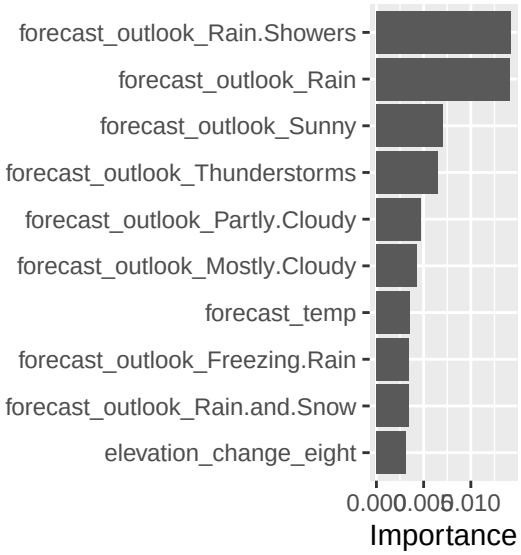
```
#> == Workflow ====================================================================
#> Preprocessor: Recipe
#> Model: boost_tree()
#>
#> -- Preprocessor ----------------------------------------------------------------
#> 3 Recipe Steps
#>
#> * step_dummy()
#> * step_zv()
#> * step_normalize()
#>
#> -- Model ----------------------------------------------------------------------
#> Boosted Tree Model Specification (regression)
#>
#> Main Arguments:
#>   trees = 100
#>   learn_rate = 0.1
#>
#> Computational engine: xgboost
```

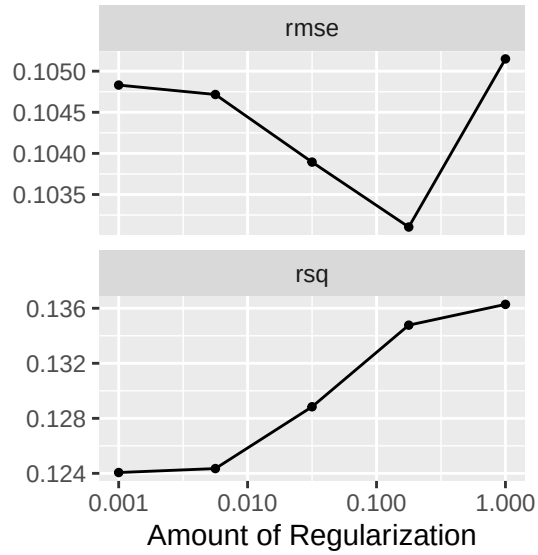



```
#> [1] 0.05014724
```

#Model refinement:

The top 5 variables are `forecast_outlook`, `observed_temp`, `forecast_temp`, `forecast_outlook_Rain.Showers`, `forecast_outlook_Sunny`. We will remove these variables from the model to see how this changes our prediction of `log_observed_precip`.



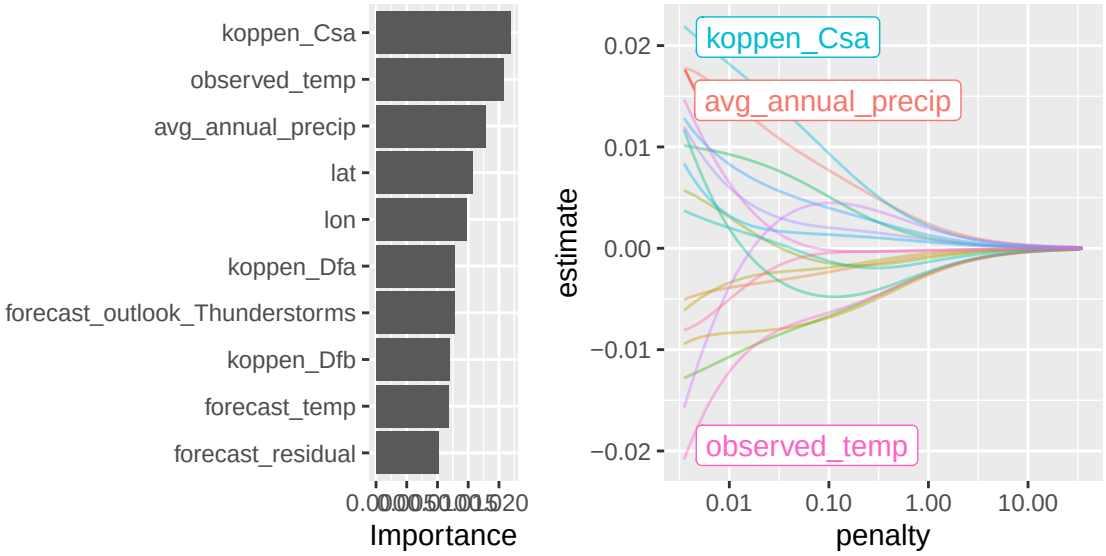



Below is our adjusted ridge model. # Adjusted Ridge Model

```
#> # A tibble: 1 x 2
#>   penalty .config
#>     <dbl> <chr>
#> 1       1 Preprocessor1_Model5
```

```
#> # A tibble: 1 x 3
#>   .metric .estimator .estimate
#>   <chr>   <chr>          <dbl>
#> 1 rsq     standard      0.0507
```

```
#> # A tibble: 18 x 3
#>    term                             estimate penalty
#>    <chr>                               <dbl>   <dbl>
#>  1 (Intercept)                       0.0660         1
#>  2 observed_temp                    -0.00246        1
#>  3 forecast_temp                    -0.00229        1
#>  4 lon                              -0.000182       1
#>  5 lat                               0.00200        1
#>  6 elevation                        -0.00264        1
#>  7 distance_to_coast                -0.000739       1
#>  8 wind                             -0.000239       1
#>  9 elevation_change_four            -0.000950       1
#> 10 elevation_change_eight           -0.000494       1
#> 11 avg_annual_precip                0.00243        1
#> 12 forecast_residual                0.00101        1
#> 13 koppen_Cfb                      -0.00135        1
#> 14 koppen_Csa                       0.00219        1
#> 15 koppen_Csb                       0.000612       1
#> 16 koppen_Dfa                       0.00131        1
#> 17 koppen_Dfb                       0.000861       1
#> 18 forecast_outlook_Thunderstorms -0.00246         1
```




In our new model we get vip with these explanatory variables, `forecast_residual`, `avg_annula_precip`, `koppen_Cfa` and `distance_to_coast`.

The rsq value for this new model is 0.0388 compared to the rsq value from the previous model of 0.0430. This model without the exlanatory variables `forecast_outlook`, `observed_temp`, `forecast_temp`, `forecast_outlook_Rain.Showers`, `forecast_outlook_Sunny is better at predicting`log_observed_precip‘.

Conclusion

- Based on our models, we've learned that temperature plays a strong role in the precipitation of a given day and that weather outlook forecasts tend to be highly accurate in predicting precipitation, as seen in thunderstorms and rain showers being considered highly important by our models.
- Temperature is a strong predictor of precipitation
- Weather forecasts are useful in predicting precipitation
- There is so much variation in precipitation between days even in a single city that the average annual precipitation for that city is largely unhelpful in predicting precipitation for a specific day
- After removing our top five predictors from our model, the residual between the forecasted and actual temperature ends of being an important factor in determining precipitation