NAME: Ashutosh Singh
PUID: 0031086892
Midterm Report

STEPS FOR MODEL FITTING:

1) **Loading the file**: The data set is loaded in R

2) **Data cleaning and pre-processing:**
   a) The columns were converted to proper format. 'DT00' was converted to numeric
   b) The year and month columns were converted to factors
   c) Dummy variables were created for all the months
   d) Rows having NA values were removed
   e) We are finally left with 296 rows. Overall 4 rows were removed from the dataset

3) **Exploratory data analysis:**
   a) Residential adjusted sales was taken as the y variable
   b) The residential sales is plotted against all the climate variables, weather variables, socio economic variables and month
   c) It was found that a majority of these variables, except for EMXT, EMNT, DT90, DP01 and VISIB are highly linearly related with residential adjusted sales.
   d) This gives us a fair idea that the most of our explanatory variables have a linear relationship with our y variable

4) **Final x variables considered:**
   EMXP, MXSD, TPCP, TSNW, EMXT, EMNT, MMXT, MMNT, MNTM, DT90, DX32, DT00, DT32, DP01, DP05, DP10, MDPT, VISIB, WDSP, MWSPD, GUST, HTDD, CLDD, LABOR, EMP, UNEMP, UNEMPRATE, PCINCOME, GSP, Dummy variables for month

5) **Model Building:**
   a) **Linear Models:**
      i) **Ordinary least squares:** The model is fitted on the above x variables and residential sales as our y. K fold cross validation was used to validate the model, k=5. The RMSE for this model was 470
      ii) **Ridge regression:** This model was built on the idea to reduce variance by introducing regularisation terms in the cost function. This model was trained on a range of shrinkage parameters – 110,125, 137, 150, 165. I came up with these values based on a trial run where I ran the model on 80% of my data set and found where I got my minimum lambda. A range of lambda as above were taken around that value (139.22). K fold cross validation(k=5), was applied and after cross validation, best lambda was found to be 125 and corresponding RMSE is 475.37
      iii) **Lasso regression:** This model was built on the idea to reduce variance by introducing regularisation terms in the cost function. This model was trained on a range of shrinkage

parameters – 6,8, 10, 15, 19, 20. I came up with these values based on a trial run where I ran the model on 80% of my data set and found where I got my minimum lambda. A range of lambda as above were taken around that value (19.37). K fold cross validation(k=5), was applied and after cross validation, best lambda was found to be 19 and corresponding RMSE is 420.91

b) Tree based models:

i) **Regression trees:** CART was run on the data set and tuned on the basis of depth of trees. Various values of depth tried were- 2, 4, 6, 8, 10. The best value was decided by doing a k fold cross validation, k=5. Best depth as found as 6 and corresponding RMSE was 550.11

ii) **Random Forest:** Since random forest are better than regression trees as they prevent overfitting. The number of trees hyperparameter was selected in the range of – 20, 30 100, 200, 300, 500, 1000, 2000, 4000, 5000. K fold cross validation was done to select the best number of trees. I finally got number of trees as 300 and RMSE as 446

iii) **Gradient Boosting Machines:** Gradient boosting machines was tried and hyperparameters were tune for number of trees and interaction depth. The range taken for both respectively was - 20, 50, 100, 200, 300, 500, 1000, 2000, 4000, 5000 and 2,4,6,8,10. The validation was done through k fold cross validation(k=5). The best value for number of trees was 200 and depth was found as 10. The corresponding RMSE was 464.70

c) Splines:

i) **Generalised additive models:** In order to account for non-linear polynomial effects in the x variables, Generative Additive Models were built. Various models with varying degree of freedoms in the x variables were tried and finally I chose a model with all linear terms except for EMXT, EMNT, DT90, DP01 and VISIB for which quadratic functions were chosen. This was validated using k fold cross validation, k=5. The RMSE for this model was 486.92.

FINAL MODEL:

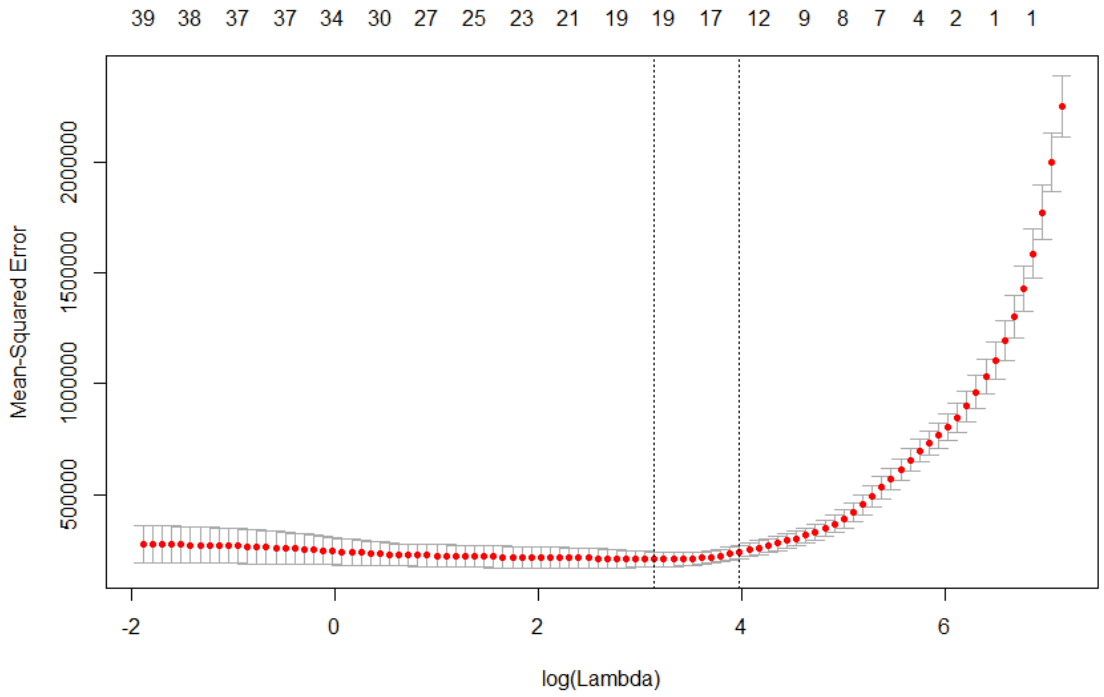
The final model chosen is the lasso regression model.

Equation for the model is as follows:

```
(Intercept)*6.515567e+03  +  EMXP*1.948217e+01
+DX32*1.623858e+03  +  DT32*-1.974503e+01  +  DP01*7.568002e+00
+DP10*6.959717e+01  +  WDSP*-1.021801e+02  +  GUST*3.159926e+01
+HTDD*2.937510e+00  +  CLDD*8.553252e+00  +  UNEMP*-1.055912e-04
+UNEMPRATE*-5.056200e+00  +month1*7.103723e+02  +month3*-3.729191e+02
+month4*-6.066368e+02  +  month5*-8.029867e+02  +  month7*4.178628e+02
+month8*7.799861e+02  +  month9*1.063747e+03  +  month10*7.746109e+02
```

```
The lambda shrinkage parameter value is taken as 19. Following is the grap
h of minimum lambda
```

Note that the final lambda was decided through k fold cross validation.



JUSTIFICATION OF THE MODEL

This model was selected on the idea that this performs the best in cross validation as compared to other linear models, tree-based models and spline-based models. This can be shown through a tabular comparison:

Model	**RMSE**
Linear Regression	470
Ridge regression	475
Lasso regression	421
CART	550
Random Forest	446
Gradient Boosting	465
GAM	487

It is evident from the above table that lasso performs the best on the cross-validation set indicating this is the best model that can be used to predict values on the unknown dataset.

"I have obeyed all rules for this exam and have not received any unauthorized aid or advice."

Ashutosh Singh