Final Presentation

Stellar Strategies(well not so much)

Executive Summary

Objective

- Our objective is to discover is there is a link between lunar and solar events and the economy

Findings

- There are some connections between solar and lunar events and the economy, but they are not strong enough to be used for predictions

Future Work

- Have some type of time based model, such as an RNN, since the economy is heavily dependent on previous days.
- Create a model that considers the effects of events on the days after the event

Overview

- Employ linear regression to assess whether astronomical events hold predictive power over market fluctuations.
- Recognize the stock market's susceptibility to various influences and acknowledge potential inaccuracies.
- Supplement the analysis with classification algorithms to categorize market movements as up, or down
- Additionally, employ feature selection techniques to identify the most effective event sets for market prediction.

Overview-Changes

- While going through the data, we had decided to do outlier cleaning, but we decided to do analysis with the outliers as well
- In the plan we did not plan on have to balance the dataset, we did not expect the big difference in the class counts
- Originally we had planned on doing a non-linear regression model, but after a closer look at the residuals from the linear model we realized there was no need

Task	Assigned To	Start Date	End Date	Week 11	12	13	14	15
Data Preparation								
Import all datasets into R	Parth&Khush	3/18	3/24					
Remove missing values	Parth	3/18	3/24					
Value Checking	Khush	3/18	3/24					
Feature Data summary	Parth	3/25	3/31					
Outlier Cleaning	Khush	3/25	3/31					
Data Transformation								
Select Relevant Attributes	Khush	3/25	3/31					
Exploring Remaining Attributes	Khush	3/25	3/31					
Standardize date formats	Parth	3/25	3/31					
Data Aggregration	Parth	3/25	3/31					
Encode Flag attribute	Parth	3/25	3/31					
Scale Features	Khush	3/25	3/31					
Data Visualization	Khush	3/25	3/31					
Correlation Martix	Parth	3/25	3/31					
Data Analysis								
Train/Test Split	Parth	4/1	4/7					
Build Linear model	Parth	4/1	4/7					
Fit Linear Model	Parth	4/1	4/7					
Evaluate Linear Model	Parth	4/1	4/7					
Build Non-linear model	Khush	4/1	4/7					
Fit Non-Linear Model	Khush	4/1	4/7					
Evaluate Non-Linear Model	Khush	4/1	4/7					
Linear/Non-Linear Testing	Parth&Khush	4/8	4/14					
Build Logistic-Regression model	Parth	4/8	4/14					
Fit Logistic-Regression Model	Parth	4/8	4/14					
Evaluate Logistic-Regression Model	Parth	4/8	4/14					
Final Report	Parth & Khush	4/15	4/21					
Final Presentation	Parth & Khush	4/15	4/21					

Data Processing

- Date standardization
  - All of the dates were encoded differently for the different datasets that we had
  - We need to standardize the date formats so that we could compare them and combine the data sets

```
flares 2010-05-01T01:34:00[1] " "
stocks 04/11/2024[1] " "
moons 15 January 1900[1] " "
lunar 2000 January 21[1] " "
solar 2000 February 5
```

```
flares "2010-05-01 01:34:00 UTC"
stocks "2024-04-11"
moons "1900-01-15"
lunar "2000-01-21"
solar "2000-02-05"
```

- Making columns with $ to numbers
  - The dataset on the stocks had everything encoded as char
  - So we had to remove the extra symbols and convert them into numeric

Data Aggregation

To aggregate the data we…

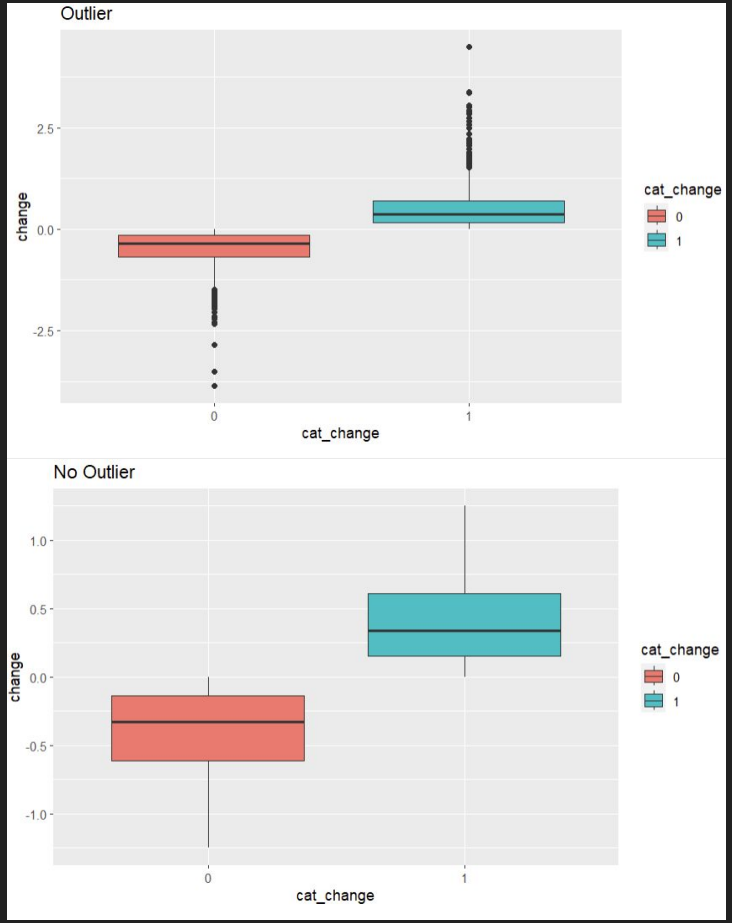
- Started with the stocks data set
- Then we added the full moon dataset by checking if the date was in the full moon dataset, and indicating that in the stocks dataset, also added the flags feature from the full moon dataset as well for matching dates
- Then added a column that showed if there was a flare on that day, by matching to see if that date was in the flares dataset
- We did that for the lunar and solar eclipses too

Feature Creation

- Creating the change variable
  - We wanted to model the change that the events cause in the economy, so we need to create a feature that had the change in prince from the opening and closing price.
- Adding the positive and negative change variable
  - We were also going to see if we could model if there would be a positive or negative change due to the event, so we created a feature that showed a positive or negative change
- Encoding the flag var
  - There are different types of full moons, which were indicated in the dataset by different type of flags.
  - To encode these we used one hot encoding so that each type of flag became a predictor.
- Has_event
  - This variable explained if there was any type of event that happened on that date

Outlier Detection

- As we observe there were many such outliers
- We removed the outliers that were present in the dataset
- We decided to keep both datasets and run analysis on both datasets, since we suspect many of those outliers could be caused by astronomical events.



Boxplots for the American Airlines data

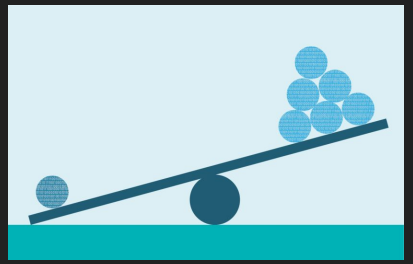
Data Exploration

Problems

- Our feature columns were all HEAVILY imbalanced
- This is because most of our cosmic events are rare and don't occur everyday
- Distributions(class0:class1)
  - Moon was 2430:86
  - Slash was 2509:7
  - Plus was 2513:3
  - Star was 2513:3
  - Two was 2508:8
  - Lunar was 2498:18
  - Solar was 2502:14
  - Flare was 2515:1
- Flare was the most significantly imbalanced

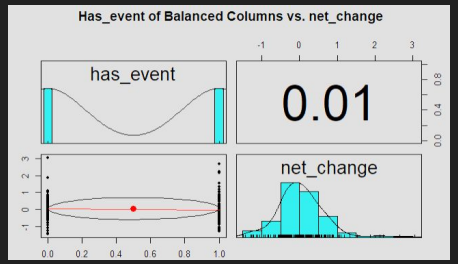
Solution(sort of)

- To address this imbalance we took the number of rows where any of the predictor columns had a minority value of 1 and sampled the same number of rows from the majority class(0)
- A lot of info was lost in this but it was necessary(as we will see later)
- The new distributions were
  - Moon was 116:86
  - Slash was 195:7
  - Plus was 199:3
  - Star was 199:3
  - Two was 194:8
  - Lunar was 184:18
  - Solar was 188:14
  - Flare was 201:1

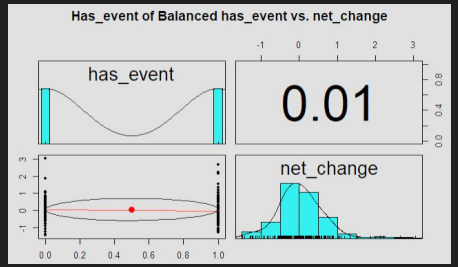


has_event variable

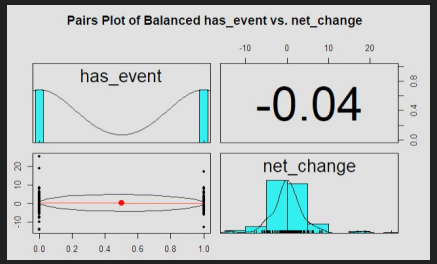
- I was noticing very faint correlation between any of the variables and the target classes.
- So I introduced the has_event variable which is 1 if ANY of the cosmic events is occurring else 0
- But I soon realized that the correlations will just cancel out and this variable will be even closer to 0 than any of the others



-AAL



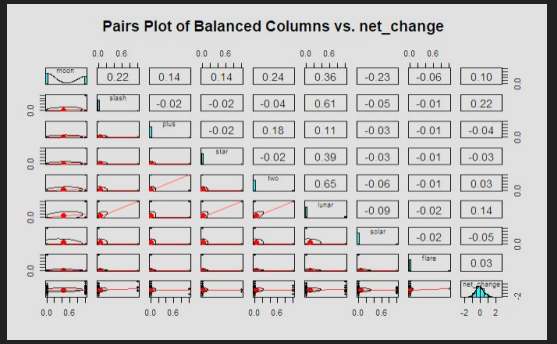
-BA

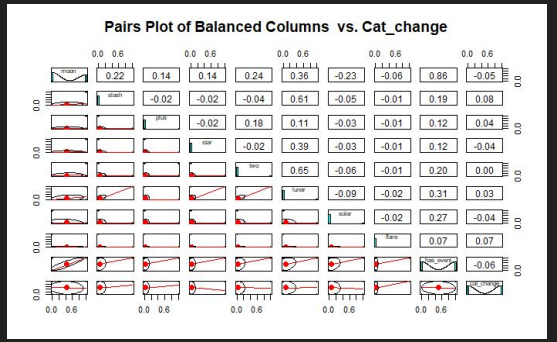


-LMT

American Airlines

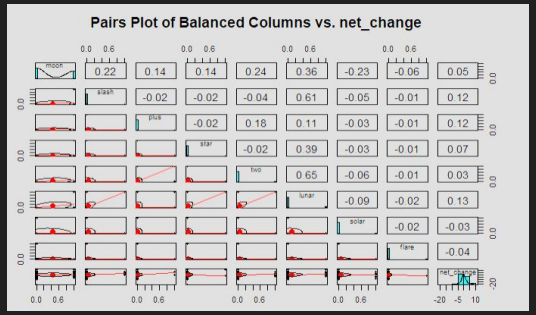
- This was the most surprising one for us because we honestly though we would see a lot more correlation here
- I tried to find correlations between the feature columns and both the net_change and the cat_change
- But the correlations were very small for both and
  - The best correlation seen was between slash and the net_change variable as 0.22
  - Similarly slash had the highest correlation when we look at cat_change but it was even lower at a mere 0.08
- This year solar eclipse(April 8, 2024)
  - net_change = +0.08
  - cat_change = 1

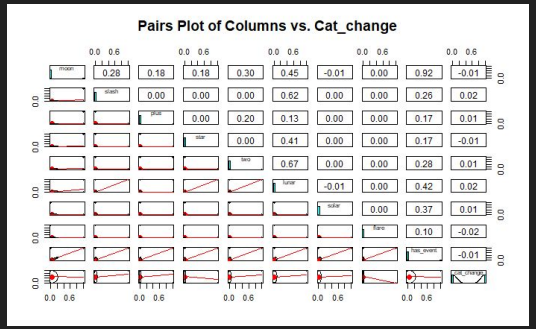




Boeing Co

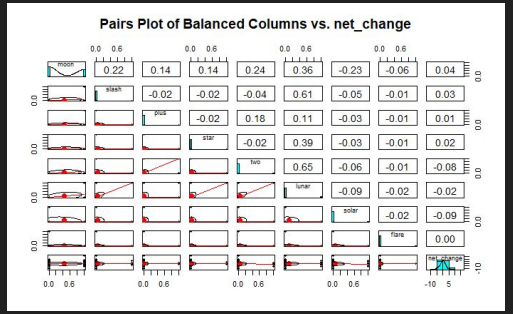
- I tried to find correlations between the feature columns and both the net_change and the cat_change
- But the correlations were very small for both
  - Slash, plus, and two had correlation coefficients greater than 0.10 vs net_change but none of the others had anything close
  - And when we looked at cat_change none of them seem to be significant at all
- This year solar eclipse(April 8, 2024)
  - net_change = -0.58
  - cat_change = 0
  - Although this could be because of their recent problems

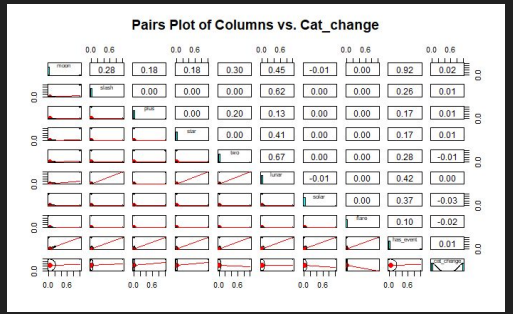




Lockheed Martin

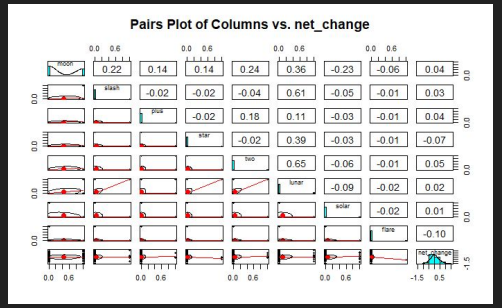
- This was not that surprising for us, and we expected a very faint correlation here
- And according to the correlation graphs we made NONE of our cosmic events had even the slightest of correlations
- The best correlation was solar vs net_change that too was a mere -0.09
- This year solar eclipse(April 8, 2024)
    - net_change = -3.62
    - cat_change = 0
    - Now this is more like the result we were hoping to see with solar eclipses and AAL

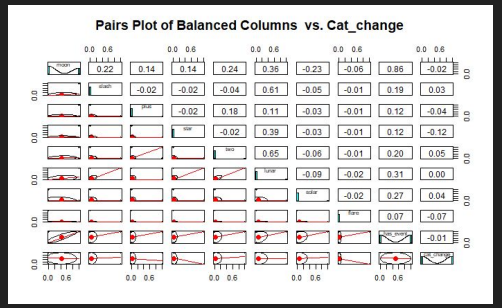




S&P500

- This we hoped would just tell us how the combination of a few companies stocks would do vs solar events
  - And we expected the correlations to be low because this is a non-volatile stock
  - But after looking at the rest of the dataset we think solar events just don't have that much of an impact to begin with
- Surprisingly this is the only one where the correlations seem to have a higher magnitude when compared with cat_change instead of net_change
- Correlations
  - Flare had the highest correlation with net_change but I would look at it with a side eye because our entire dataset has only one row where the where flare occurs
  - Star had the highest correlation with cat_change and even that was just 0.12
- This year solar eclipse(April 8, 2024)
  - net_change = -8.98
  - cat_change = 0
  - Again surprisingly this one had the highest change on the 8 April this year than any of the other stock we looked at(it is still less than .5% of a change for an index this big), even though we thought it would have the least change





Models

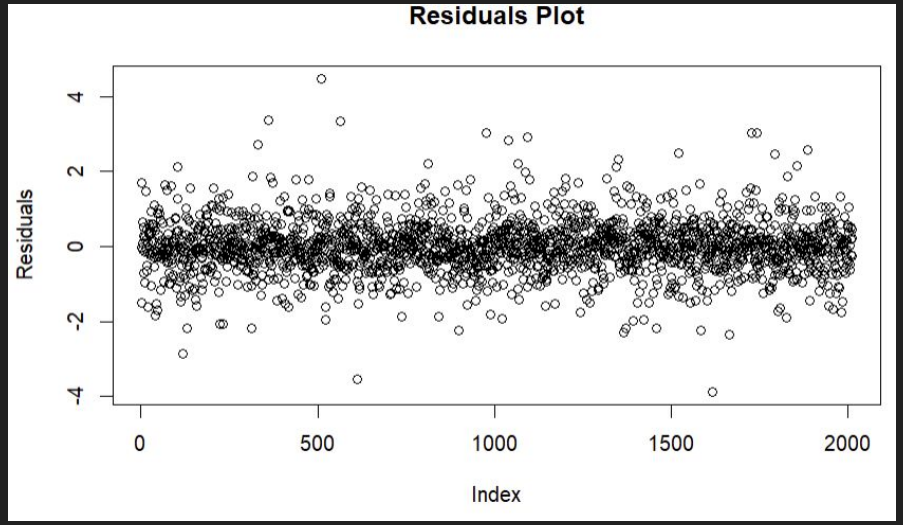
Linear Model

- Ran a simple linear regression model to predict the the change between open and close
- We see that only the slash predictor has any significance on the model for the American Airlines dataset
- The mse for this model was 0.4078255, though the mse looks small but the values for target variable are already small
- All of our other model statistics indicate that the predictors do not have any predictive power.

```
r2 0.005508158
Call:
lm(formula = change ~ moon + slash + plus + star + two + flare +
    lunar + solar, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8877 -0.3777 -0.0177  0.3623  4.4823

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.017667   0.016047   1.101  0.27105
moon        -0.009188   0.098642  -0.093  0.92580
slash        0.953521   0.330343   2.886  0.00394 **
plus        -0.288935   0.423126  -0.683  0.49478
star        -0.338479   0.508535  -0.666  0.50575
two          0.191369   0.287077   0.667  0.50510
flare        0.302333   0.706064   0.428  0.66856
lunar              NA         NA      NA       NA
solar       -0.197667   0.235841  -0.838  0.40205
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7059 on 2005 degrees of freedom
Multiple R-squared:  0.005508,  Adjusted R-squared:  0.002036
F-statistic: 1.586 on 7 and 2005 DF,  p-value: 0.1348
```

Residuals Plot

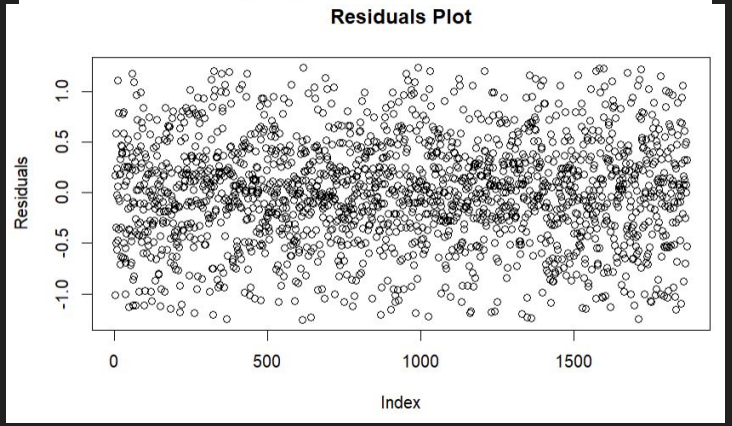
- The residual plot shows that there are some predicted values that are outliers
- Other than those outliers we see that the residuals do not have any pattern
- This indicates that there is not a need for
- non-linear regression



Linear Model without Outliers

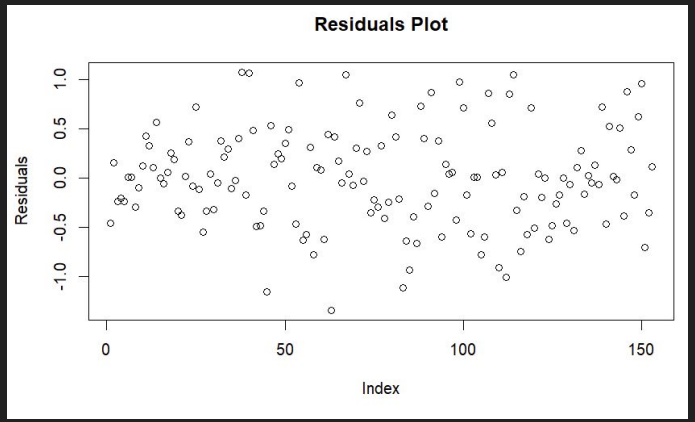
- Here we see again the slash is the only significant feature
- The mse was 0.2671165, which was better than the original model
- The rest of the model stats also indicate that this was a better model, but still not good enough to predict
- Residuals are also more scattered
- This disproves our theory that the outliers may have been caused by the events

```
Call:
lm(formula = change ~ moon + slash + plus + star + two + flare +
    lunar + solar, data = train_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.25978 -0.31978  0.00772  0.33525  1.24022

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.009782   0.012102   0.808  0.41905
moon        -0.056313   0.073011  -0.771  0.44063
slash        0.634532   0.240099   2.643  0.00829 **
plus        -0.155178   0.307167  -0.505  0.61348
star        -0.283468   0.369247  -0.768  0.44277
two         -0.044872   0.208679  -0.215  0.82977
flare        0.310218   0.512313   0.606  0.54490
lunar              NA         NA      NA       NA
solar       -0.104782   0.148345  -0.706  0.48007
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5122 on 1862 degrees of freedom
Multiple R-squared:  0.005095,  Adjusted R-squared:  0.001355
F-statistic: 1.362 on 7 and 1862 DF,  p-value: 0.2172
```



Balanced Dataset

- With the balanced dataset and without outliers, full moon feature also became a significant feature
- The mse and the rest of the model indicators were very similar to the model when we had just removed the outliers




```
Call:
lm(formula = change ~ moon + slash + plus + star + two + flare +
    lunar + solar, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-1.3422 -0.3322  0.0000  0.3186  1.0786

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.12225    0.05753   2.125   0.0353 *
moon        -0.19089    0.09191  -2.077   0.0396 *
slash        0.43464    0.23686   1.835   0.0686 .
plus        -0.13970    0.30283  -0.461   0.6453
star        -0.07136    0.30013  -0.238   0.8124
two         -0.02497    0.20587  -0.121   0.9036
flare        0.19775    0.50806   0.389   0.6977
lunar             NA         NA      NA       NA
solar       -0.27003    0.17783  -1.519   0.1311
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5048 on 145 degrees of freedom
Multiple R-squared:  0.06156,   Adjusted R-squared:  0.01625
F-statistic: 1.359 on 7 and 145 DF,  p-value: 0.2272
```

This may indicate that if we had more balanced dataset with evenly distributed classes in the features then the other features would show significance.

Penumbral Lunar Eclipse(slash)

- One thing that we noticed between the different stocks was that they were all affected by the penumbral lunar eclipse(slash) feature

```
Call:
lm(formula = change ~ moon + slash + plus + star + two + flare +
    lunar + solar, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1291 -0.4591 -0.0191  0.4309  6.2709

Coefficients: (2 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.02912    0.01812   1.607   0.1082
moon        -0.13365    0.11026  -1.212   0.2256
slash        0.73953    0.34281   2.157   0.0311 *
plus         0.04968    0.47570   0.104   0.9168
star        -1.17547    0.80371  -1.463   0.1437
two          0.16457    0.30426   0.541   0.5886
flare             NA         NA      NA       NA
lunar             NA         NA      NA       NA
solar       -0.05639    0.24078  -0.234   0.8149
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7963 on 2006 degrees of freedom
Multiple R-squared:  0.003846,  Adjusted R-squared:  0.0008668
F-statistic: 1.291 on 6 and 2006 DF,  p-value: 0.2579
```

Boeing

```
Call:
lm(formula = change ~ moon + slash + plus + star + two + flare +
    lunar + solar, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1380 -0.4580 -0.0296  0.4120  6.2620

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.03795    0.01766   2.149   0.0317 *
moon        -0.17835    0.10762  -1.657   0.0976 .
slash        0.90790    0.40277   2.254   0.0243 *
plus         0.00600    0.47053   0.013   0.9898
star        -0.27460    0.55962  -0.491   0.6237
two          0.40320    0.37104   1.087   0.2773
flare        0.19205    0.77726   0.247   0.8049
lunar             NA         NA      NA       NA
solar       -0.01795    0.24636  -0.073   0.9419
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7771 on 2005 degrees of freedom
Multiple R-squared:  0.003751,  Adjusted R-squared:  0.0002729
F-statistic: 1.078 on 7 and 2005 DF,  p-value: 0.3745
```

Lockheed Martin

```
Call:
lm(formula = change ~ moon + slash + plus + star + two + flare +
    lunar + solar, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1241 -0.4641 -0.0241  0.4177  6.2759

Coefficients: (2 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.02413    0.01765   1.367   0.1716
moon        -0.12187    0.10935  -1.114   0.2652
slash        0.81773    0.33440   2.445   0.0146 *
plus         0.04495    0.46336   0.097   0.9227
star        -0.15227    0.46044  -0.331   0.7409
two          0.15837    0.29683   0.534   0.5937
flare             NA         NA      NA       NA
lunar             NA         NA      NA       NA
solar       -0.11567    0.21575  -0.536   0.5919
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7753 on 2006 degrees of freedom
Multiple R-squared:  0.003373,  Adjusted R-squared:  0.0003925
F-statistic: 1.132 on 6 and 2006 DF,  p-value: 0.3412
```

Delta Airlines

```
Call:
lm(formula = change ~ moon + slash + plus + star + two + flare +
    lunar + solar, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9887 -0.6787 -0.0287  0.6313  6.5713

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.03866    0.02651   1.458   0.1449
moon        -0.02154    0.15994  -0.135   0.8929
slash        1.20287    0.46738   2.574   0.0101 *
plus        -0.24999    0.85838  -0.291   0.7709
star         0.54672    0.83807   0.652   0.5142
two          0.65573    0.47923   1.368   0.1714
flare        0.19134    1.16434   0.164   0.8695
lunar             NA         NA      NA       NA
solar       -0.23482    0.32393  -0.725   0.4686
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.164 on 2005 degrees of freedom
Multiple R-squared:  0.005014,  Adjusted R-squared:  0.001541
F-statistic: 1.444 on 7 and 2005 DF,  p-value: 0.1834
```

United Airlines

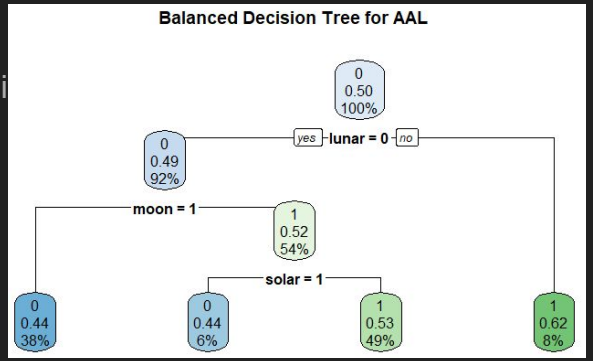
Logistic Regression

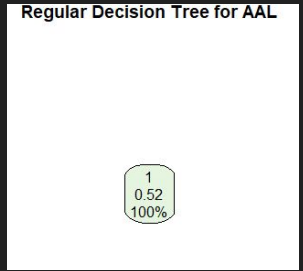
- Ran a logistic model on the balanced and the dataset without the outliers
- We predicted if there would be a positive or negative change in the stock
- The confusion matrix shows that the model learned only to predict a positive change in the data, even though the classes were balanced
- We hypothesize that this is mainly due the features not being able to carry sufficient information

```
Confusion Matrix and Statistics

          Reference
Prediction   1   0
         1 243 239
         0   9  12

               Accuracy : 0.507
                 95% CI : (0.4623, 0.5515)
    No Information Rate : 0.501
    P-Value [Acc > NIR] : 0.4118

                  Kappa : 0.0121

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.96429
            Specificity : 0.04781
         Pos Pred Value : 0.50415
         Neg Pred Value : 0.57143
             Prevalence : 0.50099
         Detection Rate : 0.48310
   Detection Prevalence : 0.95825
      Balanced Accuracy : 0.50605

       'Positive' Class : 1
```

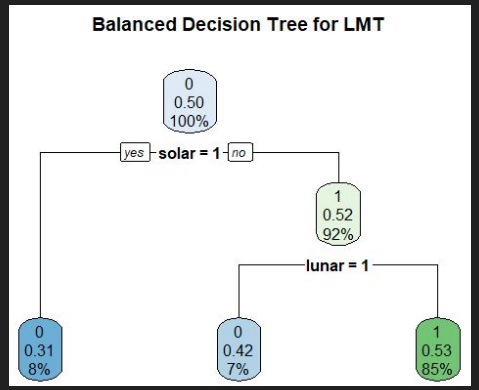
Decision Trees(Balanced vs Unbalanced)

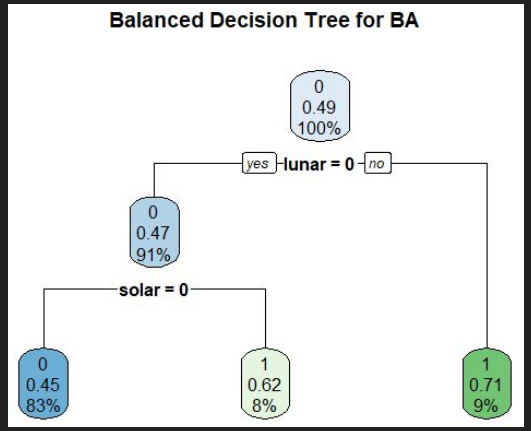
- This is where we really got to see the benefits of balancing your dataset
- It was trained on features vs cat_change
- Balanced tree seems to fit the data better
  - Rather than it just blindly predicting class 1 for cat_change it actually spli the data
- You might be wondering are there just more rows of cat_change=class1
  - And the answer is no it was a perfect split for AAL balanced data
    - We had 101 rows for both class1 and class0
  - In the unbalanced dataset there were 1214 class0 vs 1302 class1
- Metrics (for AAL test set)
  - Balanced Tree
    - Accuracy: 0.55
    - Precision: 0.65
    - Recall: 0.5416667
    - F1-score: 0.5909091
  - Unbalanced Tree
    - Accuracy: 0.4930417
    - Precision: 1
    - Recall: 0.4930417
    - F1-score: 0.6604527
- The balanced tree does better in almost all aspects but we can ignore the precision metric for unbalanced tree since it almost blindly predicts class1 and so will always have something close to precision=1 no matter what. And if we also consider that precision is part of F1 we can choose to ignore that too and hence the balanced tree performs better in all aspects. It also gives us a better understanding of which features might be most useful(in this case it would be lunar, moon, and solar)

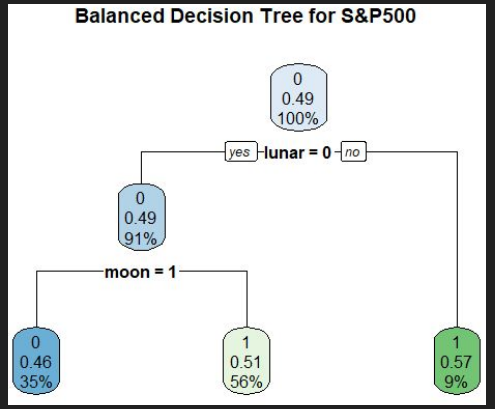


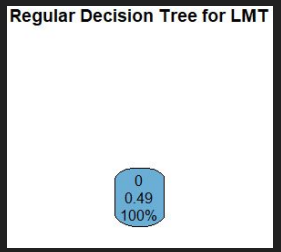


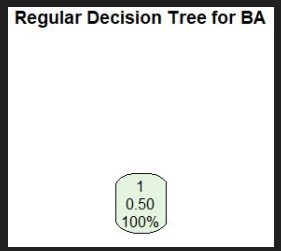
Rest of the Decision Trees

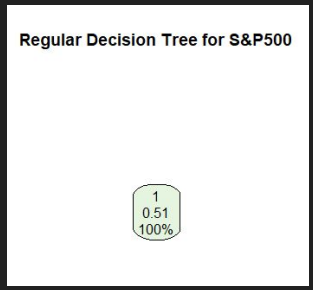






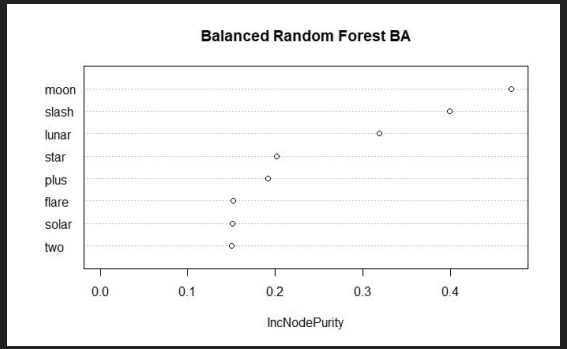


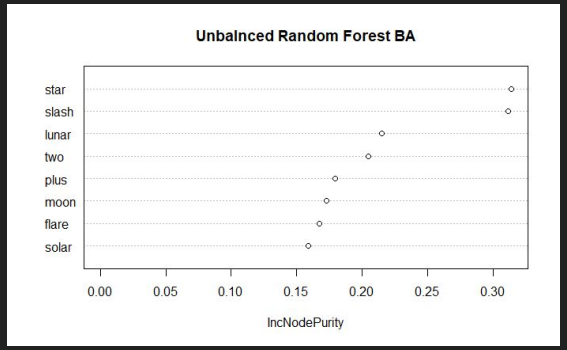




Random Forest

- Similarly we created random forests that showed similar results to decision trees
- Except in this one they both seemed to be favoring class1
- Metrics for Balanced Forest(Boeing)
  - ```
    Accuracy: 0.45
    Precision: 0.5882353
    Recall: 0.4
    F1-score: 0.4761905
              Actual
    Predicted  0  1
            0  8  7
            1 15 10
    ```
- Metrics for Unbalanced Forest
  - ```
    Accuracy: 0.5069583
    Precision: 0.9644269
    Recall: 0.505176
    F1-score: 0.6630435
              Actual
    Predicted   0   1
            0  11   9
            1 239 244
    ```
- They both seem to be favoring class1 but this is still not as apparent in the balanced model as it is in the unbalanced one
- It also shows us which features seem to have the best resulting node purity
  - For the balanced data it seems to be the moon feature
  - Whereas for the unbalanced data it seems to be the star feature





Rest of Random Forests