Business Presentation for KC Roasters

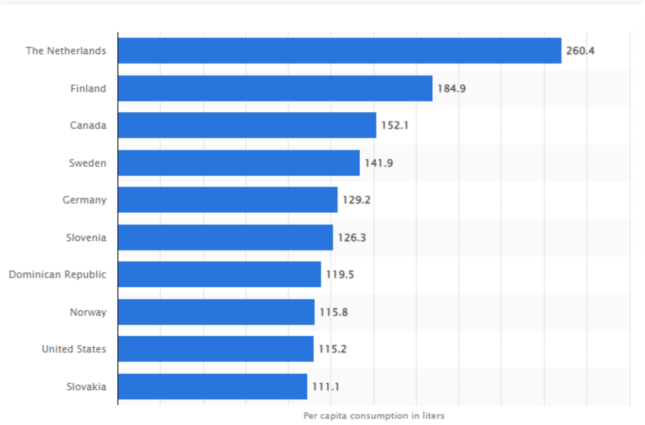
Contents

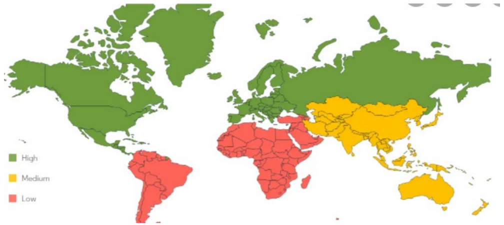
1. Overview and Business Problem
2. What can Machine Learning be expected to potentially provide?
3. Data Overview
4. EDA- Exploratory Analysis (Univariate and Bi-Variate)
5. Outliers
6. Machine Learning – Original data, OverSampled data, and Undersampling
7. Model Selection and criteria
8. Final Model and ranking of variables
9. Final EDA
10. Business recommendations and insight

KC Roasters – A Company on a quest to develop better coffee

- Coffee today has become a big business as the modern culture has accepted and in fact is consuming more coffee than ever.
- Top coffee consumers per capita rank as follows;
  - The Neatherlands- 260.4 liters per person/year
  - Finland- 184.9 liters pre person /year
  - Northern Hemisphere Countries consume the most
  - GDP for those Countries is also highest, representing a

    Opportunity for this product.





Business Problem- overview and description of the problem

- Cost of Coffee is based upon the quality of the beans after the roasting process. This process today has become highly automated and increasingly monitored by systems .
- The quality inspection process is time-consuming and a heavily manual and expensive process. Its important to determine and to reduce the quality defects in order to meet industry standards and to have a product that provides consistent deliverables to the suppliers
- Unpredictability is one of the largest business risks and challenges that KC Roasters (a wholesale coffee provider) faces..
- KC Roasters is attempting to grow in a highly dynamic and competitive market with great opportunity, however quality is extremely important since this is generally consumed by those with an increasingly discerning palette since coffee has transformed into an artisan experience with gourmet, and other high quality blends.

Business Problem- Financial Implications

- Coffee represents 1.6% of the United States total GDP , reflecting how important this industry is overall.
- Coffee quality is increasingly important to consumers as they have transitioned from casual coffee to experiencing coffee socially as an experience.
- Cost of coffee beans is getting more expensive due to
  - Demand
  - Weather problems in Brazil(worlds largest coffee producer)
  - Increased prices of coffee – roughly 24% more expensive with the rebound of the economy.
  - Therefore, production not only costs from defective product, but also lost opportunity and potentially disgruntled customers- all impacting loyalty and the brand.

Business Problem- How can Machine Learning help?

- In 2017 , only 5% of coffee producers were fully automated.
- Today this has surpassed 15% of all producers.
- Automation means systems and monitoring of every aspect of the production process is being catalogued and held in databases.
- Automation means that a more consistent product can be managed.
- The coffee bean itself has a strong relationship with how its processed to determine its quality. This represents a huge opportunity if the manufacturer can focus on those important elements to develop higher quality , which in turn will represent an increased price and demand.

Data Overview

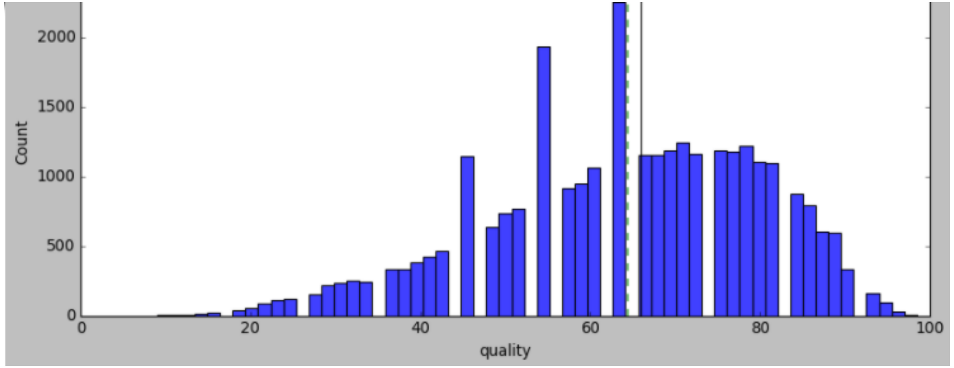
- Temperatures of different chambers of the ovens are monitored, along with the relative Humidity and the volume of raw material. A overview of the data available is below
- - T_data_1_1 - 1st sensor in the 1st chamber
- - T_data_1_2 - 2nd sensor in the 1st chamber
- - T_data_1_3 - 3rd sensor in the 1st chamber
- - T_data_2_1 - 1st sensor in the 2nd chamber
- - T_data_2_2 - 2nd sensor in the 2nd chamber
- - T_data_2_3 - 3rd sensor in the 2nd chamber
- - T_data_3_1 - 1st sensor in the 3rd chamber
- - T_data_3_2 - 2nd sensor in the 3rd chamber
- - T_data_3_3 - 3rd sensor in the 3rd chamber
- - T_data_4_1 - 1st sensor in the 4th chamber
- - T_data_4_2 - 2nd sensor in the 4th chamber
- - T_data_4_3 - 3rd sensor in the 4th chamber
- - T_data_5_1 - 1st sensor in the 5th chamber
- - T_data_5_2 - 2nd sensor in the 5th chamber
- - T_data_5_3 - 3rd sensor in the 5th chamber
- - H_data - Height of Raw material layer, basically represents the volume of raw material going inside the chamber in pounds
- - AH_data - Roasted Coffee beans relative humidity.

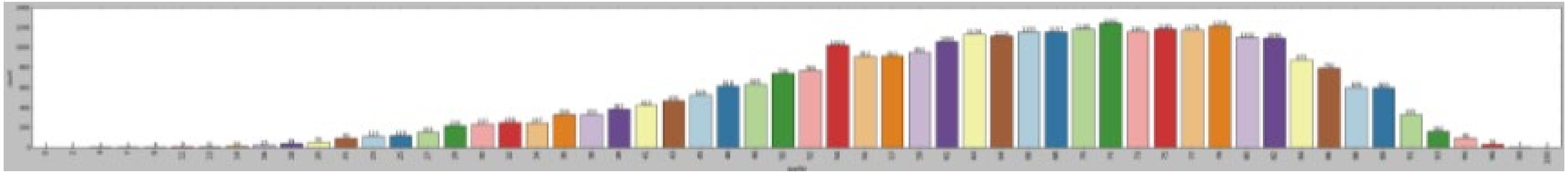
Data overview

- Data consists of 29,131 rows – (observations) with 18 columns (variables)
- Data is numeric consisting of temperature readings , humidity levels and mass
- There was some missing data, which we used median values to fill in for both H_data and AH_data. (volume of material and relative humidity)
- H_data was missing 15% of its data, AH_data was missing 11% .
- Generally, the observations provided uniform distribution in analysis

EDA- Exploratory Data Analysis

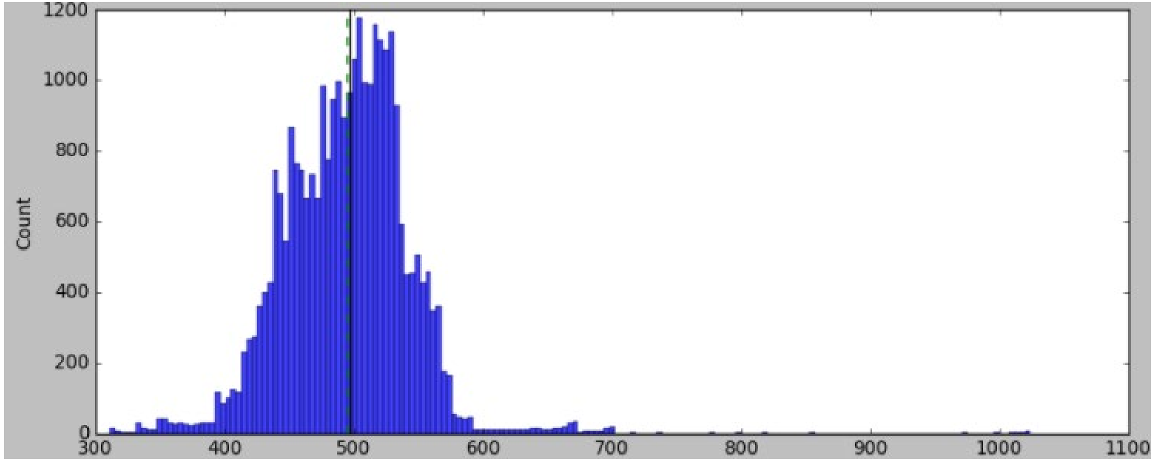
- Quality is what we are looking for – so it makes sense that we start with quality first.
    - Generally skewed in favor of higher quality, however a significant lost opportunity exists with lower quality product.
    - 50% of coffee has a rating of 66
        - 1 standard deviation= 16.397
        - 82.397-49.603 represents the IQR.





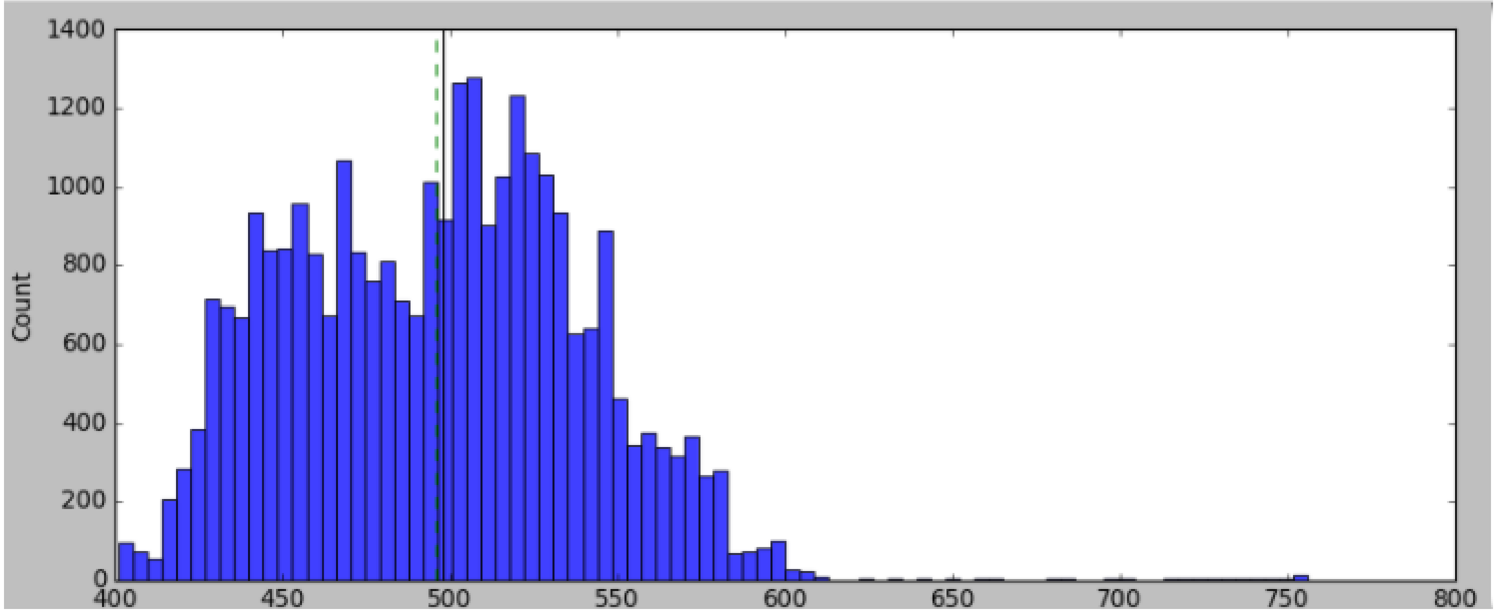
EDA- continued (Univariate)

- T_data_3_1 is the most influential variable by our analysis
- “Temperature recorded by 1st sensor in the 3rd chamber in Fahrenheit”
- Average temperature in the 3rd chamber is 494.513 degrees with a standard dev of 50.315 degrees.
- Skewed toward higher temperatures.



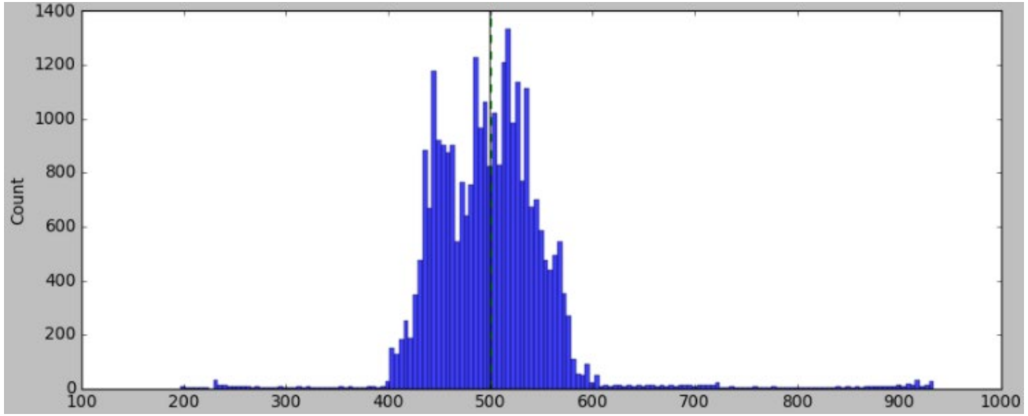
EDA continued (Univariate)

- T_data_3_2 is the 2nd most influential variable by our analysis
- 2nd sensor in the 3rd chamber



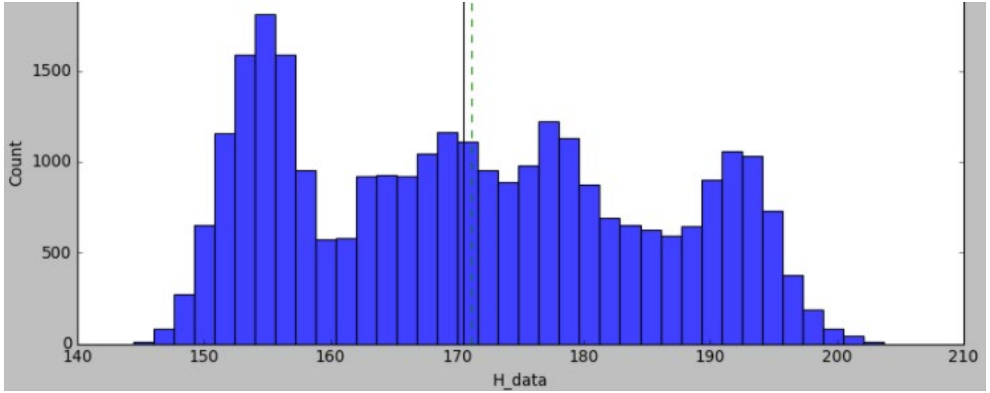
EDA continued (Univariate)

- 3rd most impactful variable is … also from the 3rd chamber T_data_3_3
- 2nd sensor in the 3rd chamber.
- This reflects the importance of the heat in the 3rd chamber



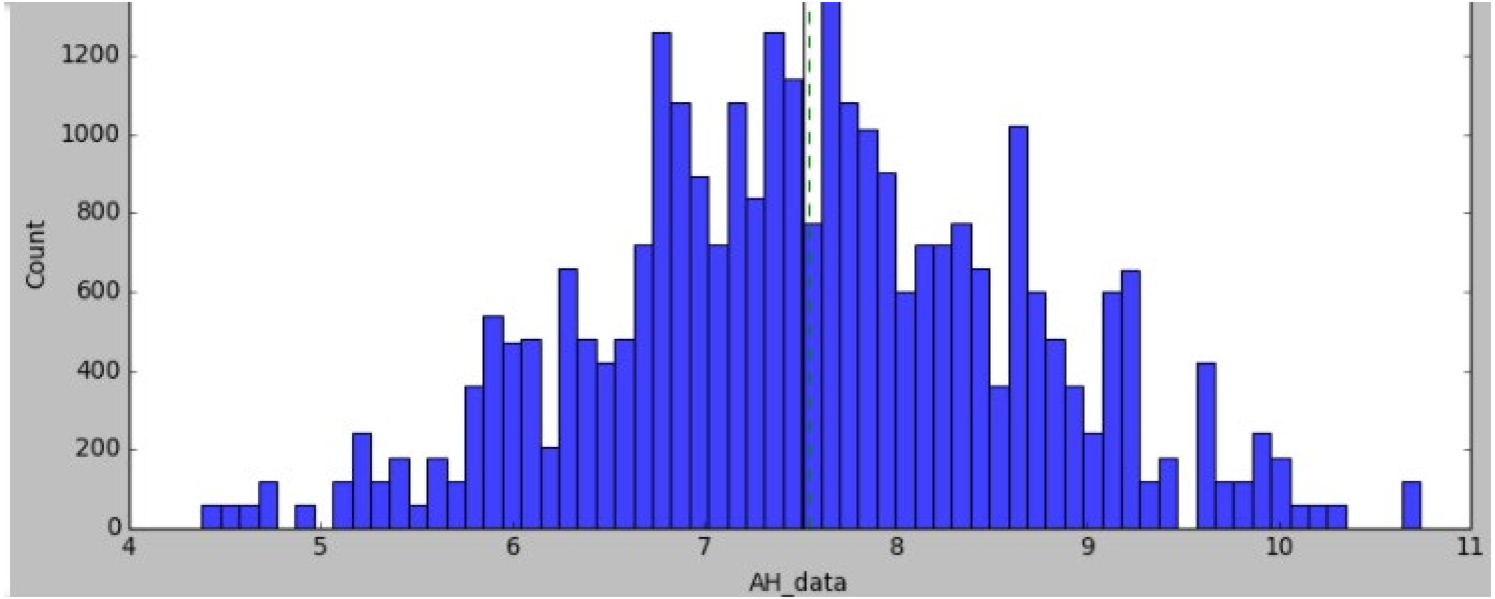
EDA- (Univariate)

- Height of Raw material layer, basically represents the volume of raw material going inside the chamber in pounds
- Outliers exist for this particular variable.
- Potential opportunity to drive consistency and potentially to determine impact if we could get more consistent in this area-



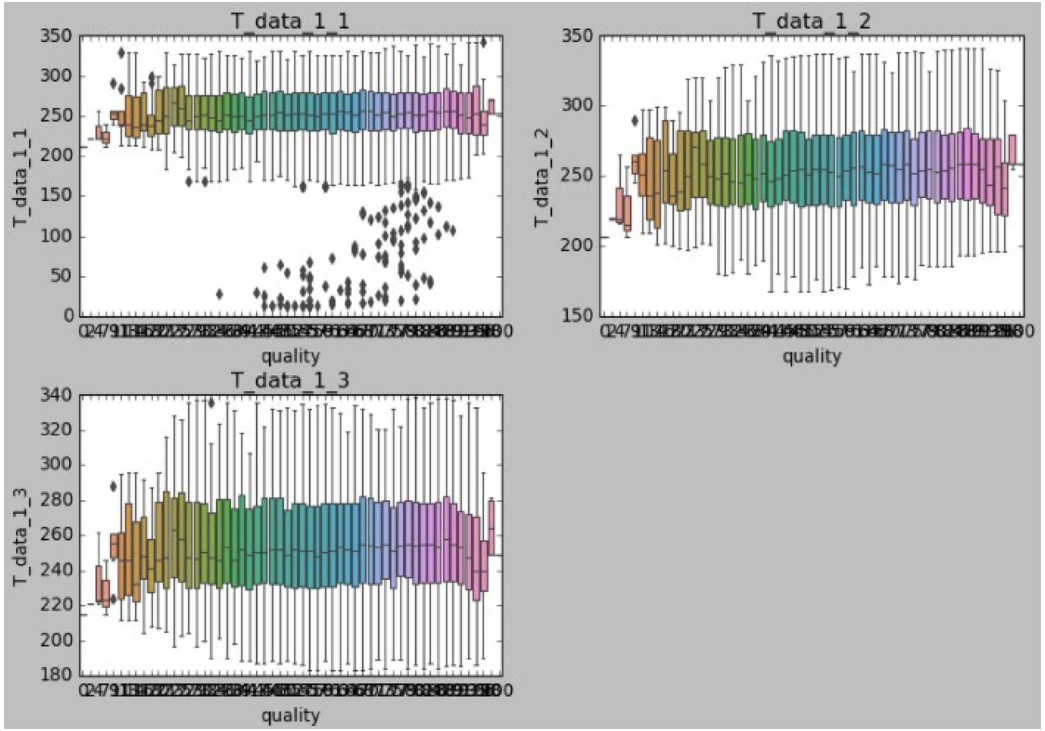
EDA- (Univariate)

- Roasted Coffee beans / relative humidity-
- Mean- 64.322 with a standard deviation of 16.397



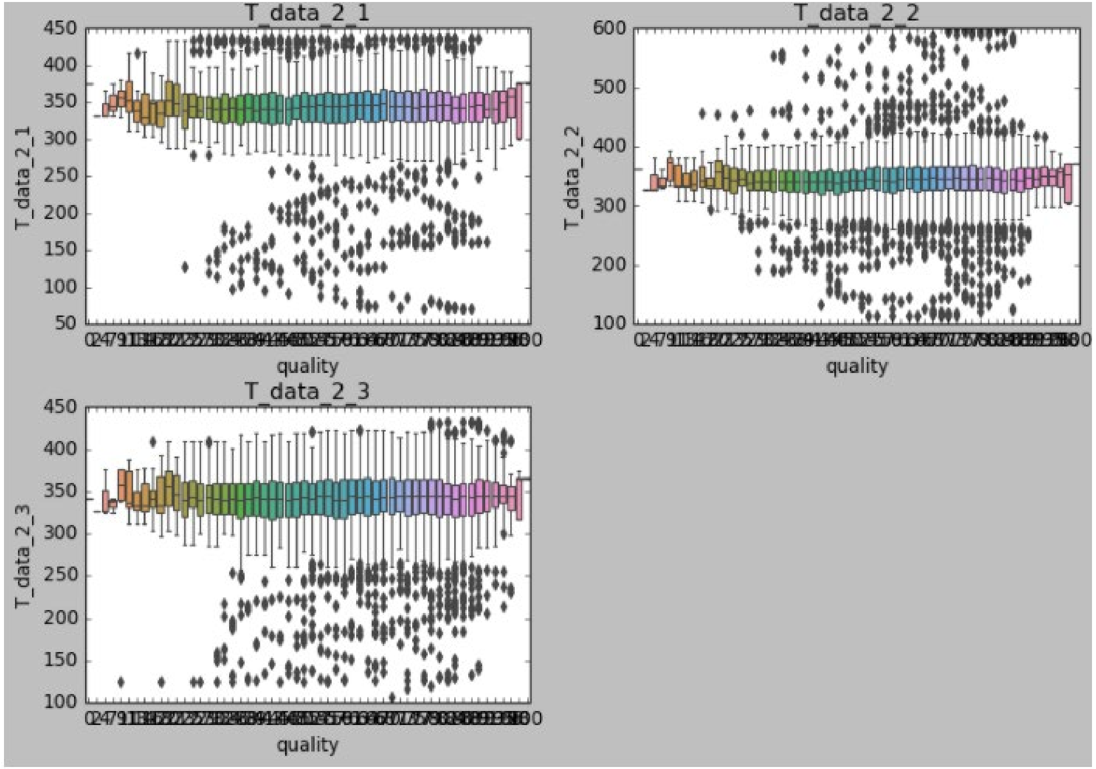
EDA- Bi-Variate analysis

- Chamber 1



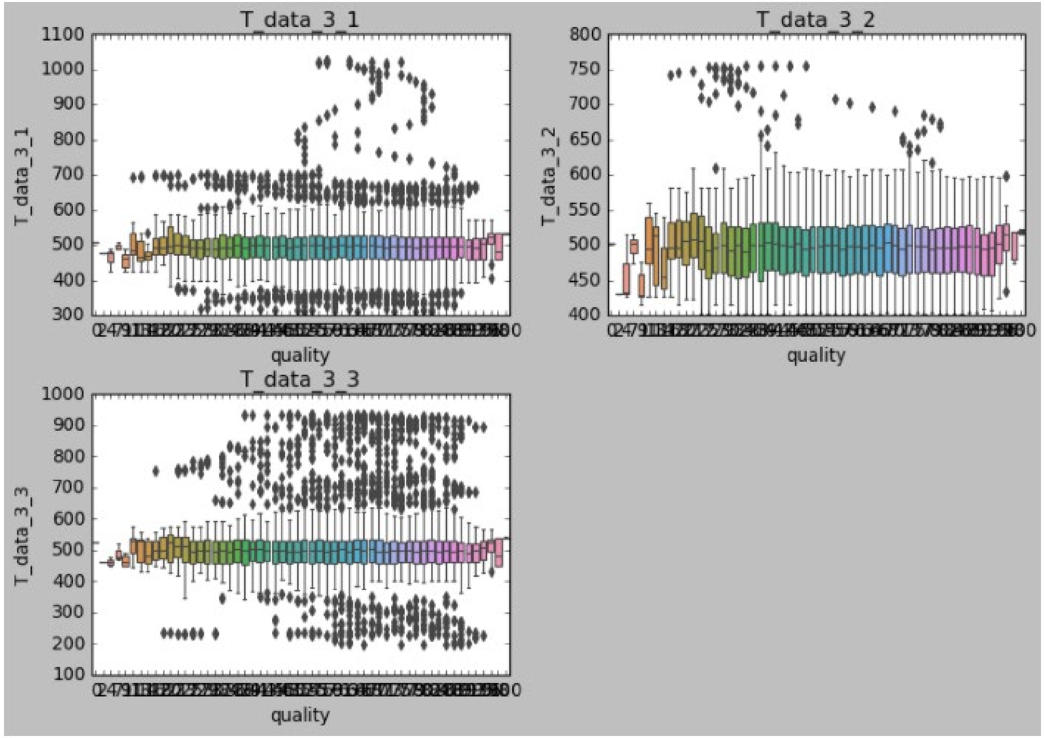
EDA- Bi-Variate analysis

- Chamber 2



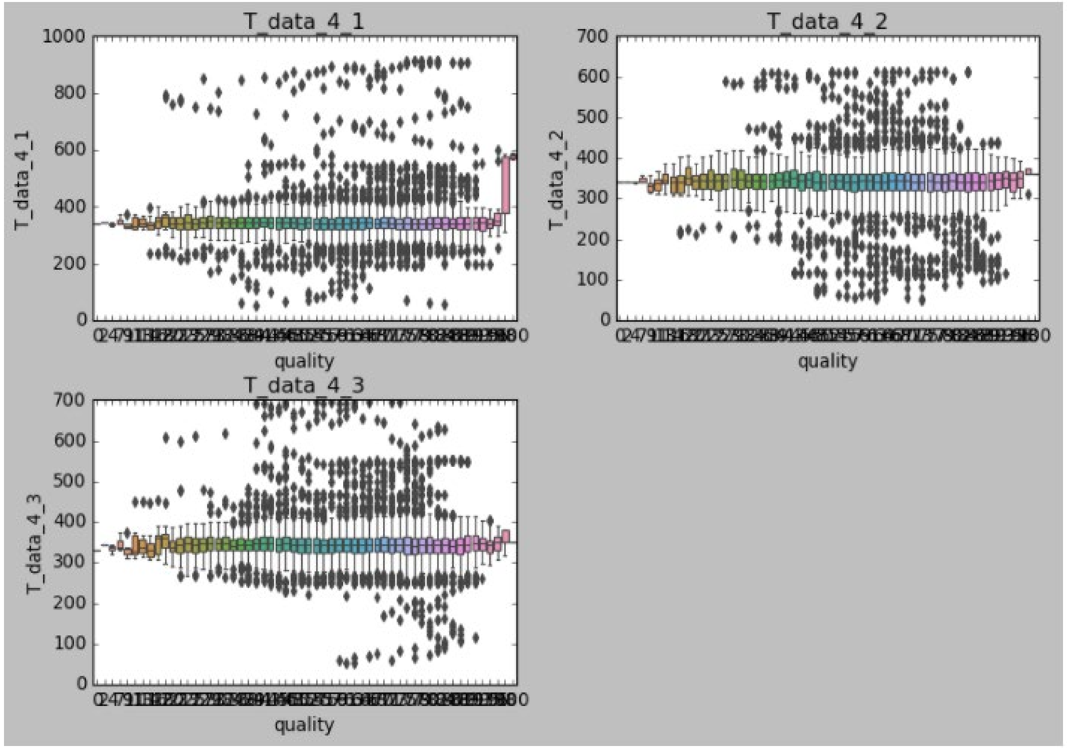
EDA- Bi-Variate analysis

- Chamber 3 vs Quality
- Notice the consistency



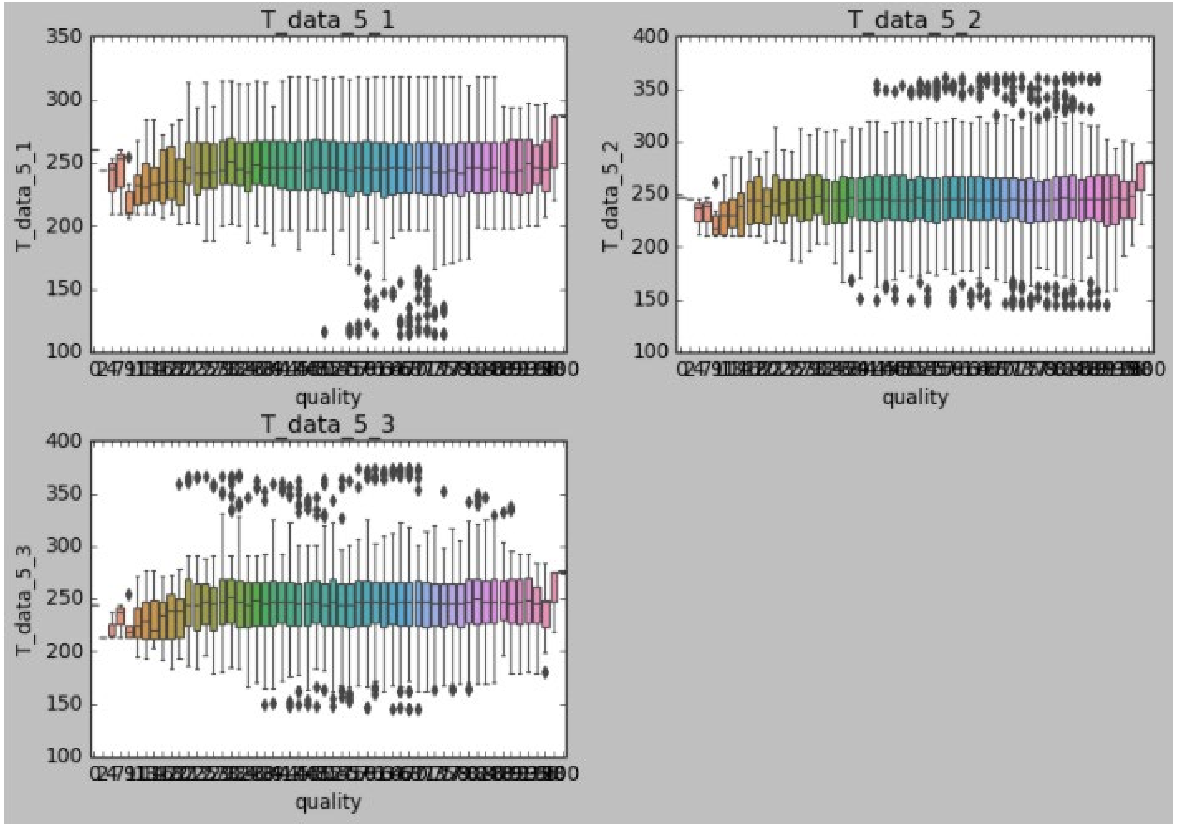
EDA- Bi-Variate analysis

- Chamber 4 vs Quality



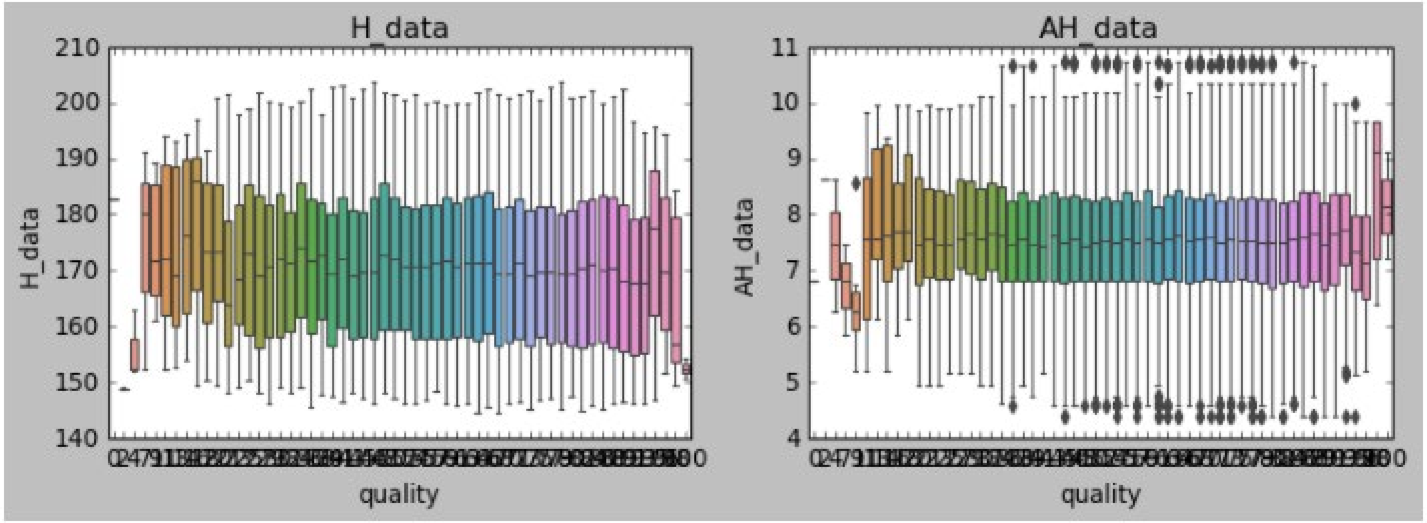
EDA- Bi-Variate analysis

- Chamber 5 vs Quality



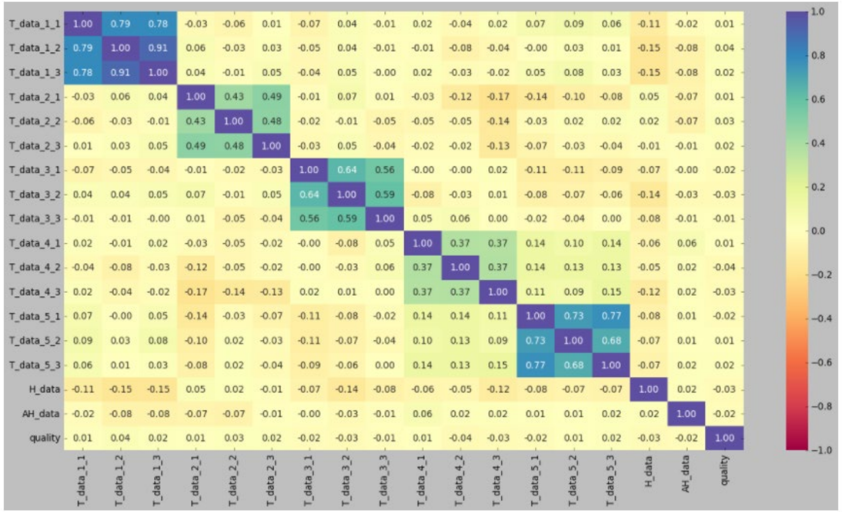
EDA- Bi-Variate analysis

- H_data vs Quality
- AH_data vs Quality

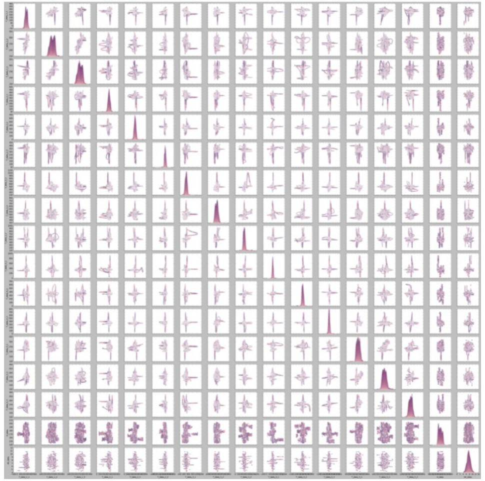


EDA- Bivariate

- Chambers 1 and 5 have the least overall impact , whereas 3 and 4 have the highest impact on quality.
- Assuming that 1 and 5 are the entry and exit points of the oven- perhaps the coffee bean requires a long, consistent approach to heat as a indicator of quality.

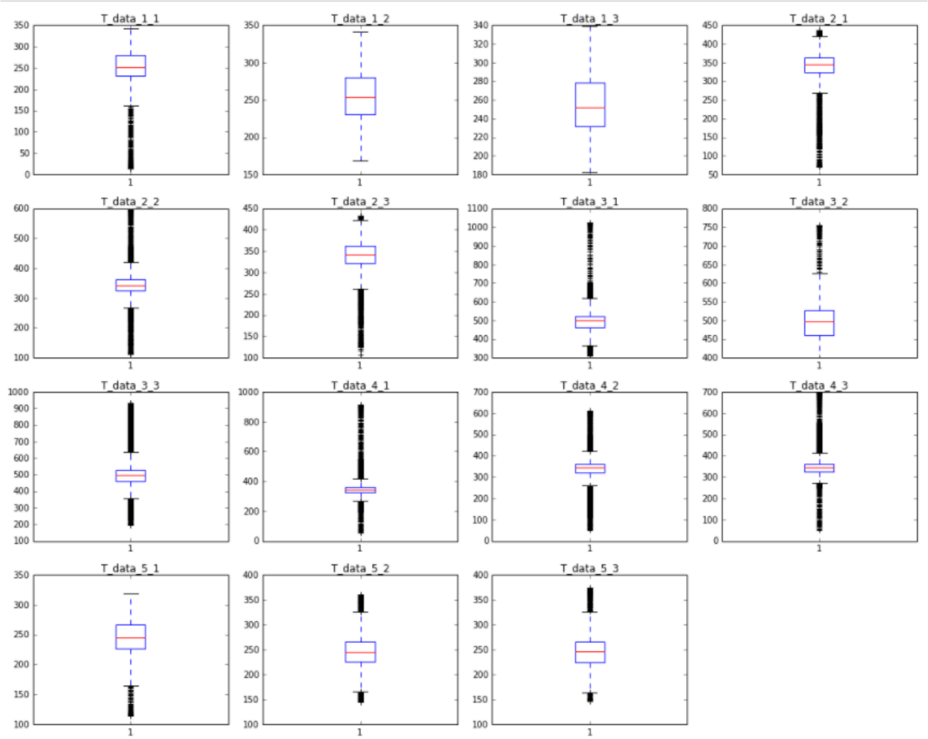


	T_data_1_1	T_data_1_2	T_data_1_3	T_data_2_1	T_data_2_2	T_data_2_3	T_data_3_1	T_data_3_2	T_data_3_3	T_data_4_1	T_data_4_2	T_data_4_3	T_data_5_1	T_data_5_2	T_data_5_3	H_data	AH_data	quality
T_data_1_1	1.00	0.79	0.78	-0.03	-0.06	0.01	-0.07	0.04	-0.01	0.02	-0.04	0.02	0.07	0.09	0.06	-0.11	-0.02	0.01
T_data_1_2	0.79	1.00	0.91	0.06	-0.03	0.03	-0.05	0.04	-0.01	-0.01	-0.08	-0.04	-0.00	0.03	0.01	-0.15	-0.08	0.04
T_data_1_3	0.78	0.91	1.00	0.04	-0.01	0.05	-0.04	0.05	-0.00	0.02	-0.03	-0.02	0.05	0.08	0.03	-0.15	-0.08	0.02
T_data_2_1	-0.03	0.06	0.04	1.00	0.43	0.49	-0.01	0.07	0.01	-0.03	-0.12	-0.17	-0.14	-0.10	-0.08	0.05	-0.07	0.01
T_data_2_2	-0.06	-0.03	-0.01	0.43	1.00	0.48	-0.02	-0.01	-0.05	-0.05	-0.05	-0.14	-0.03	0.02	0.02	0.02	-0.07	0.03
T_data_2_3	0.01	0.03	0.05	0.49	0.48	1.00	-0.03	0.05	-0.04	-0.02	-0.02	-0.13	-0.07	-0.03	-0.04	-0.01	-0.01	0.02
T_data_3_1	-0.07	-0.05	-0.04	-0.01	-0.02	-0.03	1.00	0.64	0.56	-0.00	-0.00	0.02	-0.11	-0.11	-0.09	-0.07	-0.00	-0.02
T_data_3_2	0.04	0.04	0.05	0.07	-0.01	0.05	0.64	1.00	0.59	-0.08	-0.03	0.01	-0.08	-0.07	-0.06	-0.14	-0.03	-0.03
T_data_3_3	-0.01	-0.01	-0.00	0.01	-0.05	-0.04	0.56	0.59	1.00	0.05	0.06	0.00	-0.02	-0.04	0.00	-0.08	-0.01	-0.01
T_data_4_1	0.02	-0.01	0.02	-0.03	-0.05	-0.02	-0.00	-0.08	0.05	1.00	0.37	0.37	0.14	0.10	0.14	-0.06	0.06	0.01
T_data_4_2	-0.04	-0.08	-0.03	-0.12	-0.05	-0.02	-0.00	-0.03	0.06	0.37	1.00	0.37	0.14	0.13	0.13	-0.05	0.02	-0.04
T_data_4_3	0.02	-0.04	-0.02	-0.17	-0.14	-0.13	0.02	0.01	0.00	0.37	0.37	1.00	0.11	0.09	0.15	-0.12	0.02	-0.03
T_data_5_1	0.07	-0.00	0.05	-0.14	-0.03	-0.07	-0.11	-0.08	-0.02	0.14	0.14	0.11	1.00	0.73	0.77	-0.08	0.01	-0.02
T_data_5_2	0.09	0.03	0.08	-0.10	0.02	-0.03	-0.11	-0.07	-0.04	0.10	0.13	0.09	0.73	1.00	0.68	-0.07	0.01	0.01
T_data_5_3	0.06	0.01	0.03	-0.08	0.02	-0.04	-0.09	-0.06	0.00	0.14	0.13	0.15	0.77	0.68	1.00	-0.07	0.02	0.02
H_data	-0.11	-0.15	-0.15	0.05	0.02	-0.01	-0.07	-0.14	-0.08	-0.06	-0.05	-0.12	-0.08	-0.07	-0.07	1.00	0.02	-0.03
AH_data	-0.02	-0.08	-0.08	-0.07	-0.07	-0.01	-0.00	-0.03	-0.01	0.06	0.02	0.02	0.01	0.01	0.02	0.02	1.00	-0.02
quality	0.01	0.04	0.02	0.01	0.03	0.02	-0.02	-0.03	-0.01	0.01	-0.04	-0.03	-0.02	0.01	0.02	-0.03	-0.02	1.00



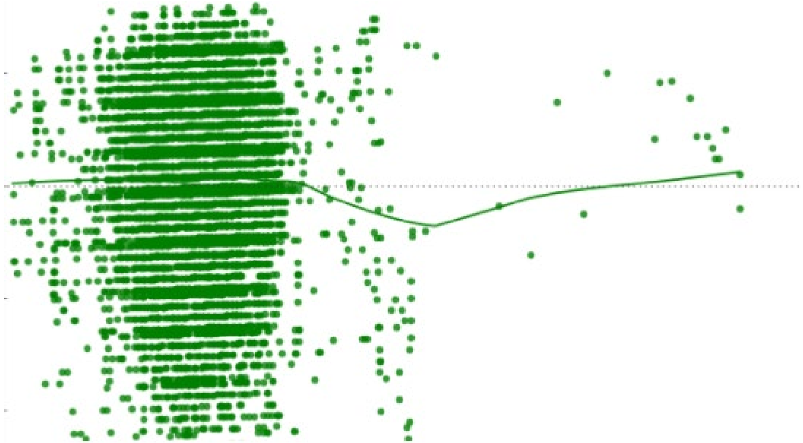
Outliers

- Outliers were present however we have models which will accompany and compensate for this .
- This was most impactful when we ran this with our linear regression model(which did poorly)



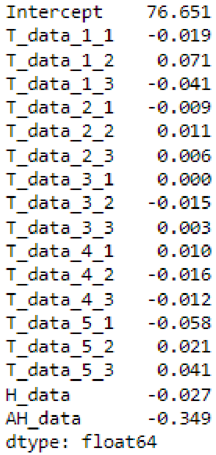
Machine Learning – models

- Linear- we ran linear separately hoping to obtain a result which could provide a predictive linear relationship to equate to a higher quality and thus higher paid coffee product. Unfortunately, the linear relationship is not very strong as displayed by the top variable to the right.

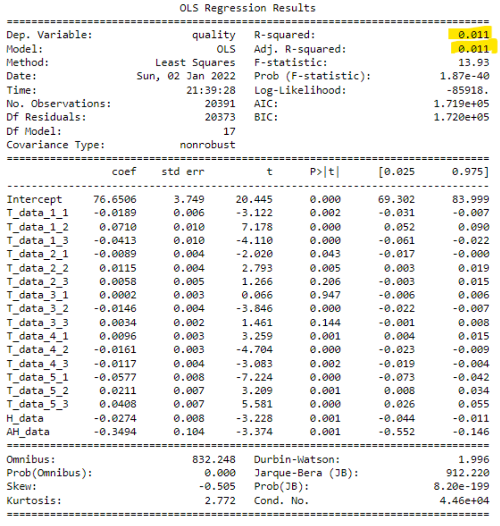


Machine Learning – models

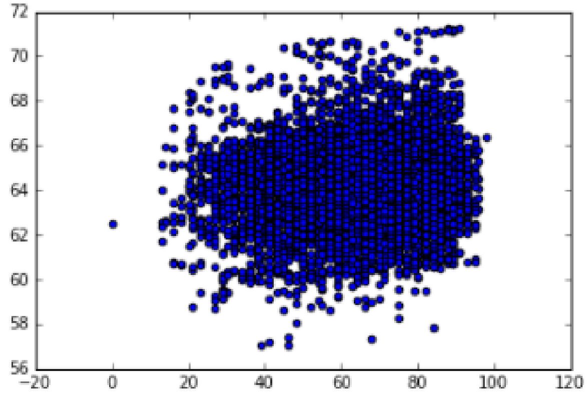
- Linear- results made this model quickly become not “material” for analysis.
- Low R-square
- Graph shows almost no linear relationship.



```
Intercept    76.651
T_data_1_1   -0.019
T_data_1_2    0.071
T_data_1_3   -0.041
T_data_2_1   -0.009
T_data_2_2    0.011
T_data_2_3    0.006
T_data_3_1    0.000
T_data_3_2   -0.015
T_data_3_3    0.003
T_data_4_1    0.010
T_data_4_2   -0.016
T_data_4_3   -0.012
T_data_5_1   -0.058
T_data_5_2    0.021
T_data_5_3    0.041
H_data       -0.027
AH_data      -0.349
dtype: float64
```



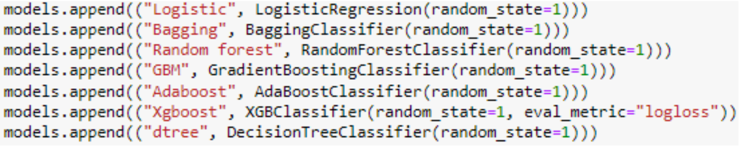
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                quality   R-squared:                       0.011
Model:                            OLS   Adj. R-squared:                  0.011
Method:                 Least Squares   F-statistic:                     13.93
Date:                Sun, 02 Jan 2022   Prob (F-statistic):           1.87e-40
Time:                        21:39:28   Log-Likelihood:                -85918.
No. Observations:               20391   AIC:                         1.719e+05
Df Residuals:                   20373   BIC:                         1.720e+05
Df Model:                          17
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     76.6506      3.749     20.445      0.000      69.302      83.999
T_data_1_1    -0.0189      0.006     -3.122      0.002      -0.031      -0.007
T_data_1_2     0.0710      0.010      7.178      0.000       0.052       0.090
T_data_1_3    -0.0413      0.010     -4.110      0.000      -0.061      -0.022
T_data_2_1    -0.0089      0.004     -2.020      0.043      -0.017      -0.000
T_data_2_2     0.0115      0.004      2.793      0.005       0.003       0.019
T_data_2_3     0.0058      0.005      1.266      0.206      -0.003       0.015
T_data_3_1     0.0002      0.003      0.066      0.947      -0.006       0.006
T_data_3_2    -0.0146      0.004     -3.846      0.000      -0.022      -0.007
T_data_3_3     0.0034      0.002      1.461      0.144      -0.001       0.008
T_data_4_1     0.0096      0.003      3.259      0.001       0.004       0.015
T_data_4_2    -0.0161      0.003     -4.704      0.000      -0.023      -0.009
T_data_4_3    -0.0117      0.004     -3.083      0.002      -0.019      -0.004
T_data_5_1    -0.0577      0.008     -7.224      0.000      -0.073      -0.042
T_data_5_2     0.0211      0.007      3.209      0.001       0.008       0.034
T_data_5_3     0.0408      0.007      5.581      0.000       0.026       0.055
H_data        -0.0274      0.008     -3.228      0.001      -0.044      -0.011
AH_data       -0.3494      0.104     -3.374      0.001      -0.552      -0.146
==============================================================================
Omnibus:                      832.248   Durbin-Watson:                   1.996
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              912.220
Skew:                          -0.505   Prob(JB):                    8.20e-199
Kurtosis:                       2.772   Cond. No.                     4.46e+04
==============================================================================
```



Machine learning – Models

- For the largest portion of analysis after the failure of the linear regression we focused on

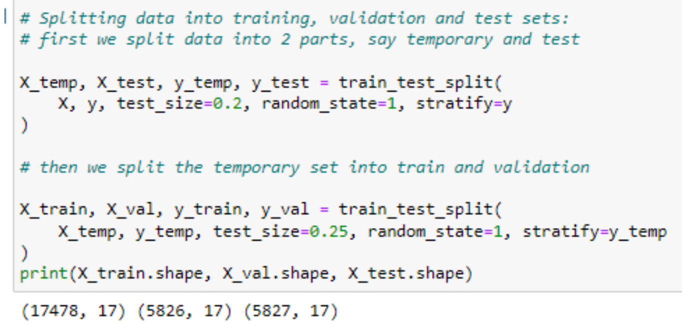
1. Logistic
2. Bagging
3. Random Forest
4. GBM- Gradient Boosting
5. Adaboost
6. XGBoost
7. Decision Tree



```
models.append(("Logistic", LogisticRegression(random_state=1)))
models.append(("Bagging", BaggingClassifier(random_state=1)))
models.append(("Random forest", RandomForestClassifier(random_state=1)))
models.append(("GBM", GradientBoostingClassifier(random_state=1)))
models.append(("Adaboost", AdaBoostClassifier(random_state=1)))
models.append(("Xgboost", XGBClassifier(random_state=1, eval_metric="logloss"))
models.append(("dtree", DecisionTreeClassifier(random_state=1)))
```

We separated the data into

1. Testing Data to develop our initial model
2. Validation Data to test our initial models
3. Testing Data to keep for the final model for production

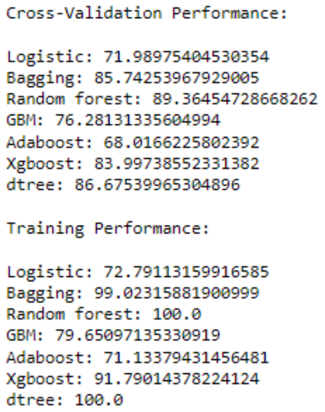


```
# Splitting data into training, validation and test sets:
# first we split data into 2 parts, say temporary and test

X_temp, X_test, y_temp, y_test = train_test_split(
    X, y, test_size=0.2, random_state=1, stratify=y
)

# then we split the temporary set into train and validation

X_train, X_val, y_train, y_val = train_test_split(
    X_temp, y_temp, test_size=0.25, random_state=1, stratify=y_temp
)
print(X_train.shape, X_val.shape, X_test.shape)
```

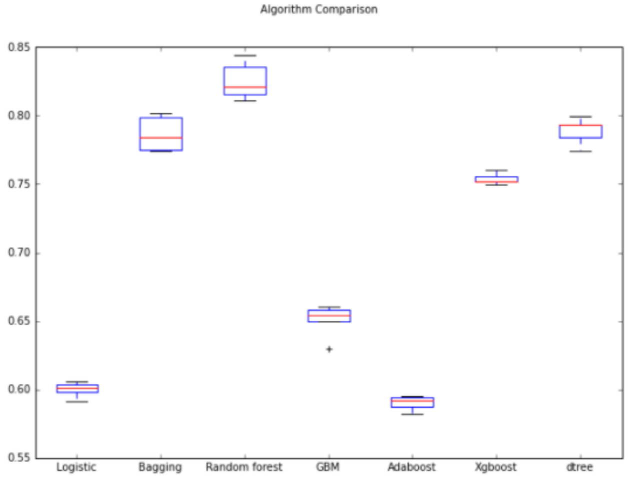
```
(17478, 17) (5826, 17) (5827, 17)
```

Machine Learning – Original Data

- We ran all models against the original Data – with results as shown after cross validation
- Bagging, Random Forest and Decision tree all were overfitting the data
- GBM, Adaboost, Logistic had the most consistent performance
  - ○ If we didn't go further, we would have chosen GBM as it provided a more consistent model than Xgboost, even though Xgboost performed better – there was a larger variance.

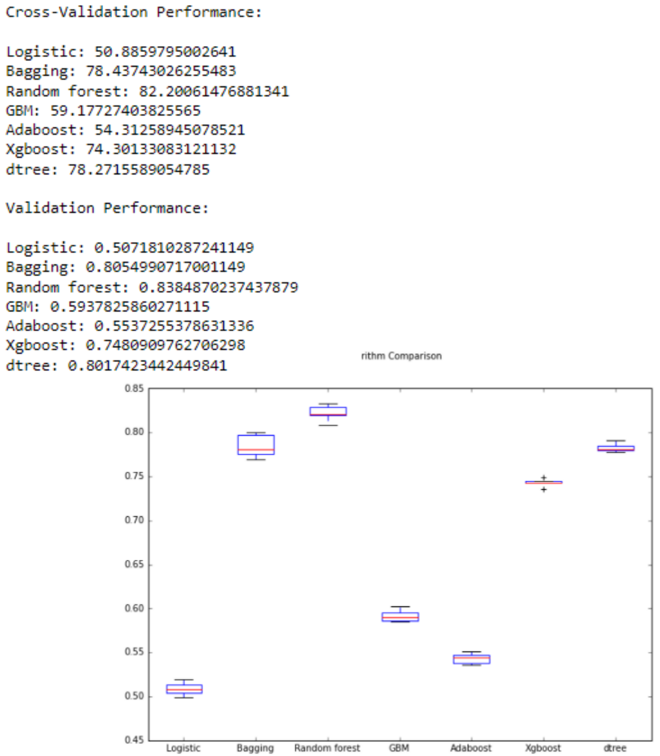


```
Cross-Validation Performance:

Logistic: 71.98975404530354
Bagging: 85.74253967929005
Random forest: 89.36454728668262
GBM: 76.28131335604994
Adaboost: 68.0166225802392
Xgboost: 83.99738552331382
dtree: 86.67539965304896

Training Performance:

Logistic: 72.79113159916585
Bagging: 99.02315881900999
Random forest: 100.0
GBM: 79.65097135330919
Adaboost: 71.13379431456481
Xgboost: 91.79014378224124
dtree: 100.0
```



Machine Learning – Oversampled Data (SMOTE)

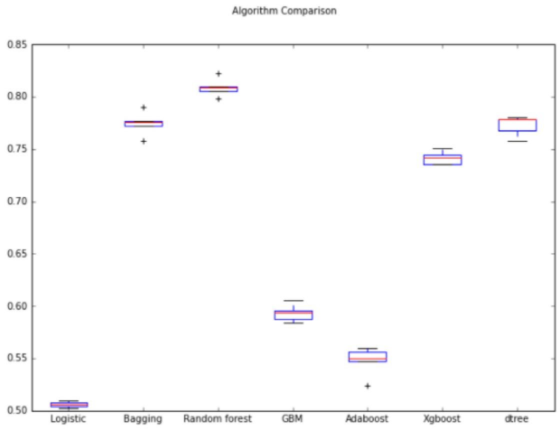
- Oversampling provided better results
- Bagging, Random forest and Decision Tree did well
- Logistic, GBM, Adaboost did poorly
- Clearly helped with overfitting of data

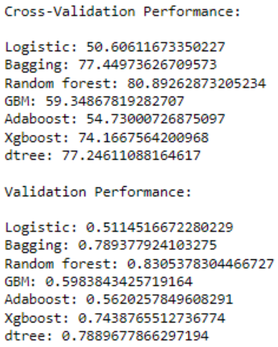


```
Cross-Validation Performance:

Logistic: 50.8859795002641
Bagging: 78.43743026255483
Random forest: 82.20061476881341
GBM: 59.17727403825565
Adaboost: 54.31258945078521
Xgboost: 74.30133083121132
dtree: 78.2715589054785

Validation Performance:

Logistic: 0.5071810287241149
Bagging: 0.8054990717001149
Random forest: 0.8384870237437879
GBM: 0.5937825860271115
Adaboost: 0.5537255378631336
Xgboost: 0.7480909762706298
dtree: 0.8017423442449841
```

Machine Learning – Undersampled data

- Undersampling similarly provided better results
- Similar to the results from oversampling
- Bagging, Random Forest, Xgboost, and Decision Tree performed better here than the other models





```
Cross-Validation Performance:

Logistic: 50.60611673350227
Bagging: 77.44973626709573
Random forest: 80.89262873205234
GBM: 59.34867819282707
Adaboost: 54.73000726875097
Xgboost: 74.1667564200968
dtree: 77.24611088164617

Validation Performance:

Logistic: 0.5114516672280229
Bagging: 0.789377924103275
Random forest: 0.8305378304466727
GBM: 0.5983843425719164
Adaboost: 0.5620257849608291
Xgboost: 0.7438765512736774
dtree: 0.7889677866297194
```

Model selection – or playoff

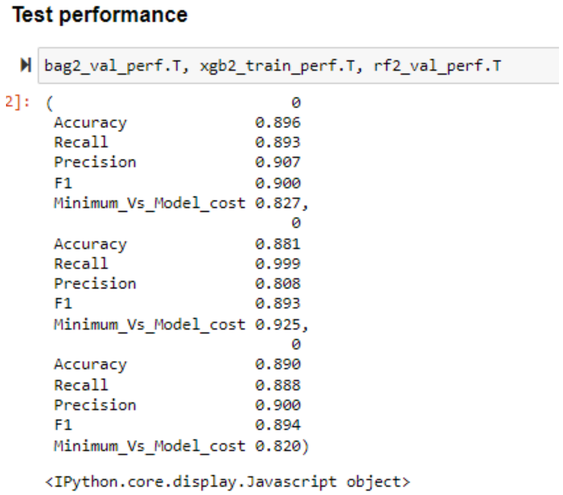
- Ultimately, we chose 3 models to compare for our final analysis

- 1. Bagging- Bagging represented well in both Oversample and Undersample. The consistency and lack of volatility made it stand out. We chose to go with the Oversampled data due to a 1 point difference in the validation performance (77-78)
- 2. XGBoost- Also a consistent performer. XGBoost did well in all three runs. It ran almost exactly the same for under and over sampled data and provided extremely consistent returns of 74 . In the Original data run, it overfitted slightly but still provided a good model overall
- Random Forest- Random Forest similar to the previous 2 examples had a consistent return on under and over sampling. Random forest had better scores

- Honorable mention goes to Decision Tree. We chose to go with the other performers but could have put Decision tree over XGBoost.

Final Model Selected

- We selected Random Forest. It was a consistent performer, in areas where it did not perform as the best model, it still performed well and consistently.
- We gave great weight to consistency.
- Random Forest was beat by both XGBoost in precision , however it consistently preformed well in Accuracy (very close) , Recall and F1 and overall model performance.



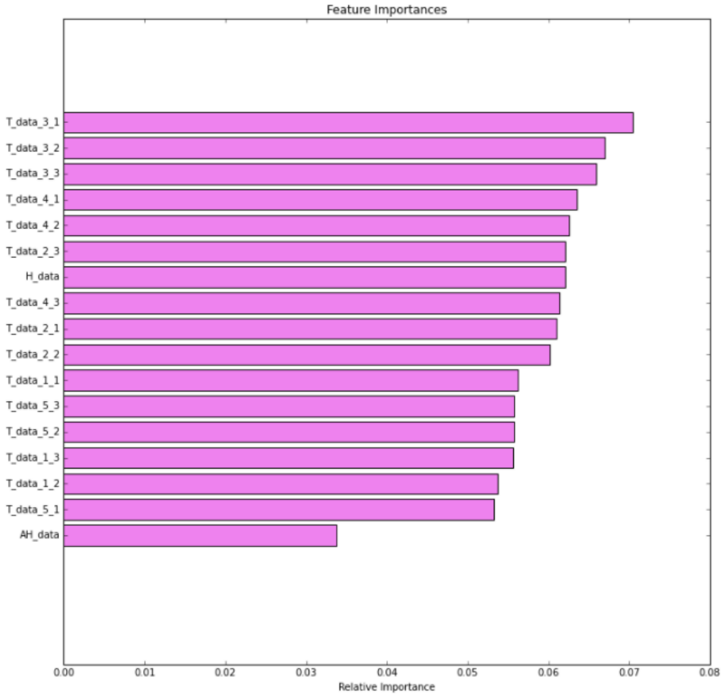
Test performance

```
bag2_val_perf.T, xgb2_train_perf.T, rf2_val_perf.T
```

```
2]: (                        0
     Accuracy             0.896
     Recall               0.893
     Precision            0.907
     F1                   0.900
     Minimum_Vs_Model_cost 0.827,
                             0
     Accuracy             0.881
     Recall               0.999
     Precision            0.808
     F1                   0.893
     Minimum_Vs_Model_cost 0.925,
                             0
     Accuracy             0.890
     Recall               0.888
     Precision            0.900
     F1                   0.894
     Minimum_Vs_Model_cost 0.820)

    <IPython.core.display.Javascript object>
```

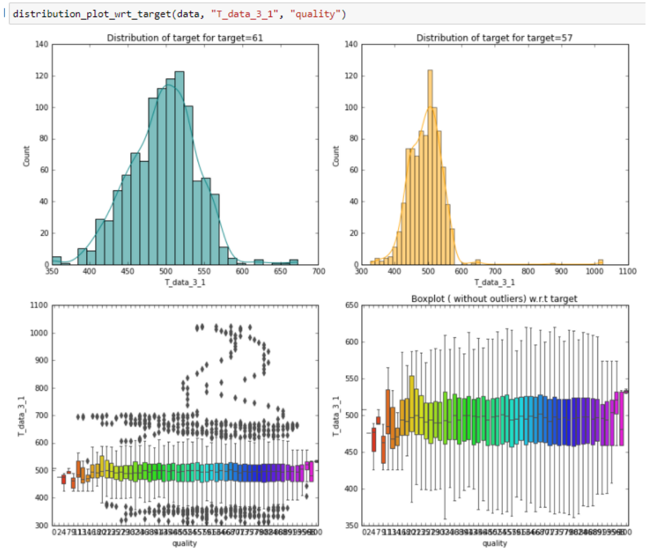
Ranking of features – as run with Random Forest

- As we indicated earlier, it appears that the middle of the oven is a very important part of the process in developing a quality bean with a flavor that demands a higher quality rating .
- Volume was near the top , even though the consistency was not very good in this area.
- Relative humidity was the least impactful.



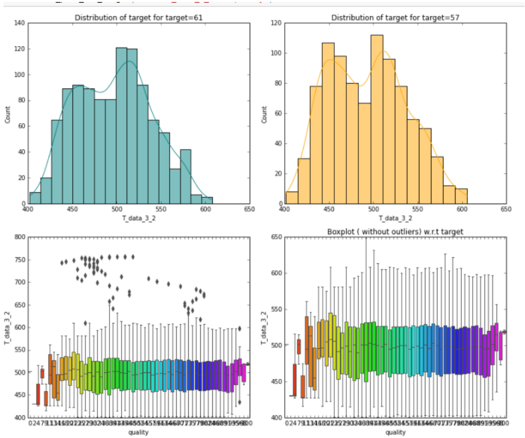
Final EDA – distribution of T_data_3_1, T_data_3_2 and T_data_3_1

- Most impactful variable
- Distribution is consistent
- Quality relationship seems clear

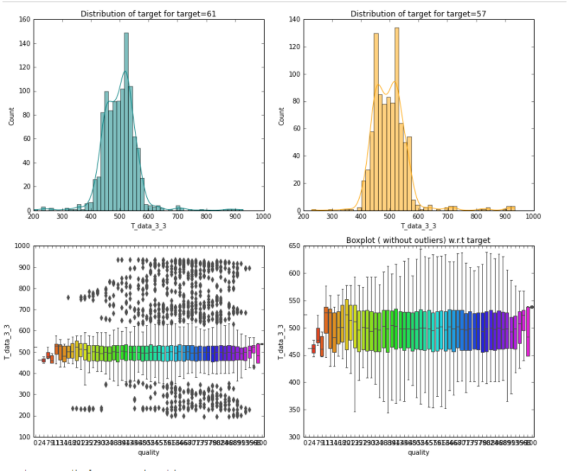


```
distribution_plot_wrt_target(data, "T_data_3_1", "quality")
```

Distribution of target for target=61
Count
T_data_3_1
Distribution of target for target=57
Count
T_data_3_1
T_data_3_1
quality
Boxplot (without outliers) w.r.t target
T_data_3_1
quality



Distribution of target for target=61
Count
T_data_3_2
Distribution of target for target=57
Count
T_data_3_2
T_data_3_2
quality
Boxplot (without outliers) w.r.t target
T_data_3_2
quality



Distribution of target for target=61
Count
T_data_3_3
Distribution of target for target=57
Count
T_data_3_3
T_data_3_3
quality
Boxplot (without outliers) w.r.t target
T_data_3_3
quality

Business Insights and Recommendations

It seems the data shows that the consistency in the ovens is very important. Its important to keep the ovens consistently running and paying particular attention to the middle of the oven- ie. 3rd chamber, 4th chamber and 5th (in that order)

- Keep chamber 3-1, 3-2 and 3-3 around 494 to 495 degrees
- Keep chamber 4-1,4-2 and 4-3 as close to 345 degrees as you can
- Keep chamber 5-1, 5-2, and 5-3 close to 245 degrees

Even though the data was not as consistent with the mass (H_data) , it did show into one of the top variables. I would recommend that the operations team attempts to do a better job to have a consistent flow of material through the oven. If we could have about a mass between 157- 185 and focus on perhaps putting minimums and maximums on the runs, it may provide valuable insight and may become more impactful. It would make sense that a consistent volume going through the ovens would provide a more consistently roasted product – but at the minimum it would provide a better variable for further analysis.

IF these recommendations can be managed, it can be expected to have a higher quality product and continued environment for increased revenues.

Thank you .

- By - Frederick Duff