Data Science Case Study

By Simon Cox

09 November 2020

1 Table of Contents

2 Executive Summary

This case study uses anonymised data sourced from an anonymous travel website and attempts to find the patterns and predictors of consumer behaviour based on the dataset provided.

A description of the data is given below:

row_id	Unique identifier for each row
search_date	The day the search was made (higher value is more recent; the date has been cast to days since an epoch)
Stage_1/2/3/4	Binary flag indicating that the user reached the first/second/third/fourth stage of the purchase funnel
Search_Feature_1/2/3/4/5	Features that relate to the search that the user conducted on our platform e.g. the dates of travel, details of who is travelling etc.
Product_Feature_1/2/3/4	Features that relate to the product that was selected by the user e.g. price of the package, board basis (all-inclusive, room only etc.)

The approach to the analysis was to:

1) Tidy the data so that it could be analysed appropriately by imputing missing values based on a number of assumptions.
2) Look for correlations between variables to understand if any instances of multicollinearity exist.
3) Look for patterns in the data within individual distributions and between variables using various visualisations.
4) Transform the data to enable meaningful analysis. For example, categorical variables were transformed into binary variables.
5) Follow a sequence of steps that looked at the collection of inputs (search date, stages and search features) against each of the outputs (product features) separately:
    a) Generate a series of models of the data
    b) Determine the feature importance of the most accurate models
    c) . Note that model accuracy is not the focus of the case study; it is the identification of the predictors of consumer behaviour.

The results for each of the product features are shown in the table below:

	Important Predictors (Ranked 1 -3)		
Product Feature	**1**	**2**	**3**
product_feature_1	search_feature_1	search_feature_5	search_feature_3
product_feature_2	search_feature_4	search_feature_3	search_date
product_feature_3	search_feature_1	stage_3	search_feature_2
product_feature_4	search_feature_1	stage_3	search_date

3 Method

3.1 Data Preparation

The CSV dataset was read into R, and an initial examination of the data was undertaken:

```
     row_id         search_date       stage_1       stage_2         stage_3         stage_4        search_feature_1
 Min.   :     0   Min.   : 0.00   Min.   :1   Min.   :0.00    Min.   :0.00    Min.   :0.00    Min.   :0.00
 1st Qu.: 75948   1st Qu.:15.00   1st Qu.:1   1st Qu.:0.00    1st Qu.:1.00    1st Qu.:0.00    1st Qu.:0.00
 Median :151896   Median :25.00   Median :1   Median :0.00    Median :1.00    Median :0.00    Median :2.00
 Mean   :151896   Mean   :23.82   Mean   :1   Mean   :0.08    Mean   :0.89    Mean   :0.02    Mean   :2.21
 3rd Qu.:227844   3rd Qu.:33.00   3rd Qu.:1   3rd Qu.:0.00    3rd Qu.:1.00    3rd Qu.:0.00    3rd Qu.:4.00
 Max.   :303792   Max.   :42.00   Max.   :1   Max.   :1.00    Max.   :1.00    Max.   :1.00    Max.   :5.00
                                              NA's   :44849   NA's   :44849   NA's   :44849
 search_feature_2 search_feature_3 search_feature_4 search_feature_5 product_feature_1 product_feature_2
 Min.   :-18.00   Min.   :  0.00   A:196836         : 45080          Min.   : 0.00     Min.   :0.000
 1st Qu.: 21.00   1st Qu.: 50.00   C:106957         A: 91298         1st Qu.: 8.00     1st Qu.:0.000
 Median : 43.00   Median : 70.00                    B:167415         Median :29.00     Median :0.000
 Mean   : 67.11   Mean   : 72.22                                     Mean   :28.65     Mean   :1.248
 3rd Qu.: 77.00   3rd Qu.: 90.00                                     3rd Qu.:39.00     3rd Qu.:2.000
 Max.   :522.00   Max.   :340.00                                     Max.   :67.00     Max.   :5.000

 product_feature_3 product_feature_4
 Min.   : 0.00     Min.   : 0.00
 1st Qu.: 4.00     1st Qu.: 4.00
 Median :10.00     Median :10.00
 Mean   :16.17     Mean   :16.87
 3rd Qu.:34.00     3rd Qu.:34.00
 Max.   :38.00     Max.   :39.00
```

This basic information presented the following findings and actions:

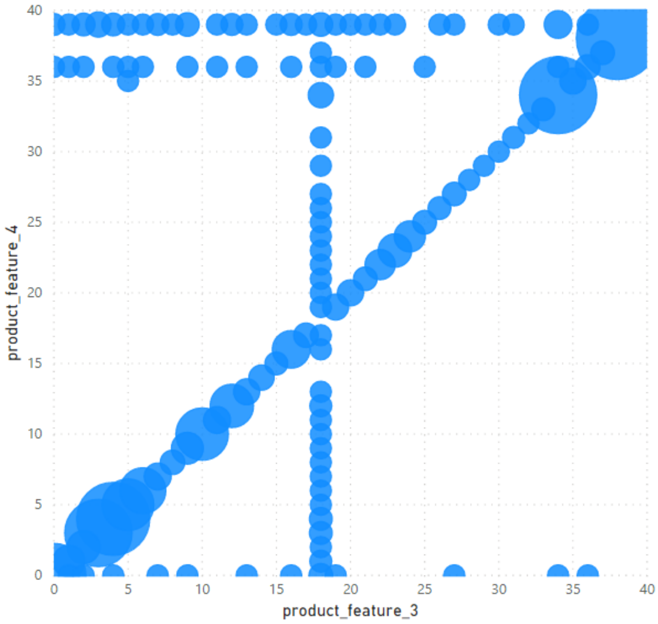
- row_id represents a sequential number and provides no predictive capability to any of the product features. Therefore this field can be omitted from the working dataset.
- stage_1 contains a value of ‘1’ for all records in the dataset. This provides no predictive capability to any of the product features and so this field can be omitted from the working dataset.
- stage_2, stage_3 and stage_4 are all missing data in 44,849 rows. Given these are all binary variables, it was assumed that a record with missing data is equivalent to a record with a value of ‘0’ (i.e. the consumer did not reach this stage of the purchase funnel). Therefore missing values for any of these fields can be imputed as having a value of 0.
    - Note that this assumption should be validated with the business, but this is not an available option for this case study.
- search_feature_5 is missing data in 45,080 rows. Given this is a categorical variable, a new value of “None” can be assigned whenever there is missing data.

3.2 Exploratory Data Analysis

3.2.1 Scatterplots

A key step is to look for any signs of relationships between the independent variables (i.e. search_date, stage_2/3/4, search_feature_1/2/3/4/5; herein referred to as “features”), and the dependent variables (i.e. product_feature_1/2/3/4, herein referred to as “labels”). It is also important to look for relationships within the set of features and within the set of labels.

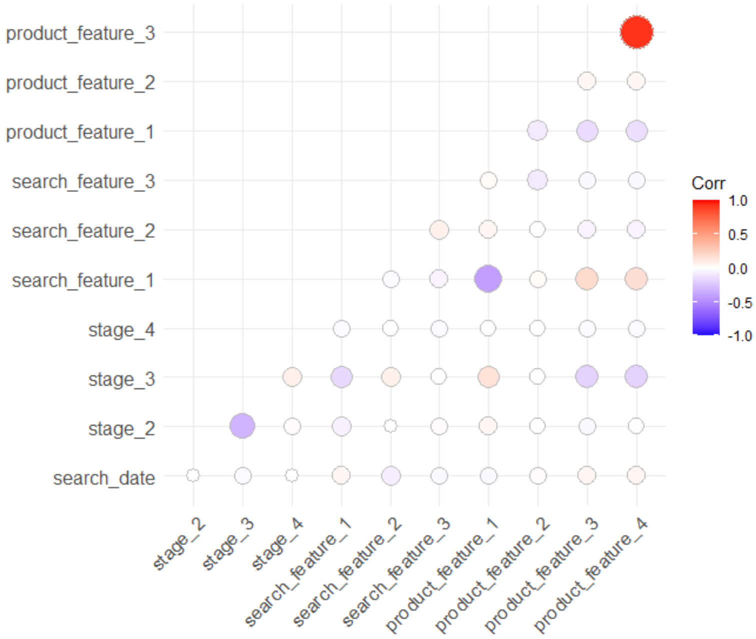
To do this a series of pairwise scatterplots were created with the dataset. Most of these plots did not show any obvious relationships, except for the following relationships between labels:



- It could be argued that there is somewhat of a positive correlation between product_feature_3 and product_feature_4 in the first plot (light blue). The sizes of the points indicate the number of corresponding records, and it is obvious that for many of the records the value of product_feature_3 is the same as for product_feature_4. Interestingly there are a couple of other characteristics of this plot:
    - There are many records where product_feature_3 is 18 but product_feature_4 can be any value. This doesn't happen for any other value of product_feature_3.
    - There are many records where product_feature_4 is either zero, 36 or 39, but product_feature_3 can be any other value. Again this doesn't happen for any other value of product_feature_4.

3.2.2 Correlation

The next step is to measure the correlation between the numeric features and labels. A correlation matrix plot is provided below:

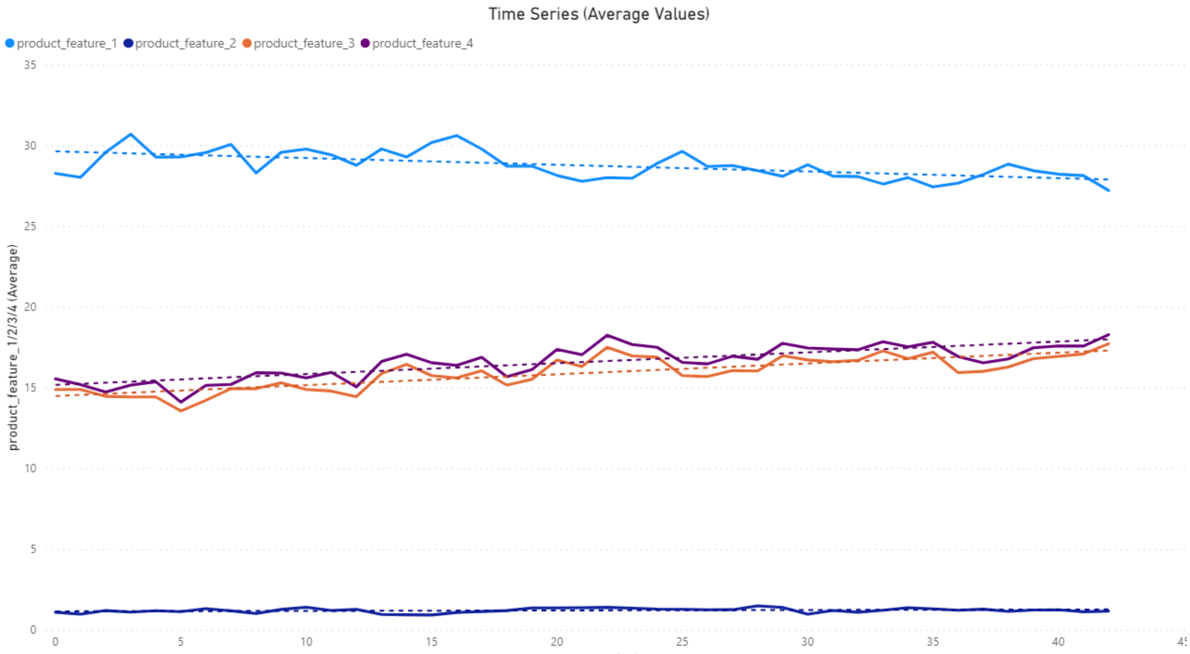


In accordance with the scatterplots, the only pair of variables with correlation of any significance is product_feature_3 and product_feature_4. No other significant correlations exist between numeric variables. This means that there is no multicollinearity between features to consider and therefore we can happily include all numeric features in any models that we build.

3.2.3 Time Series

We know that the search_date feature represents the number of days since an important event occurred. In this way it acts as an axis for a time series chart, which allows us to see if there are any trends in terms of consumer behaviour over time.

The following chart shows the average values for each of the labels over time:



- There is a clear similarity between the average values selected for product_feature_3 and product_feature_4 over time, with a slight upwards trend for both of these series.
- There is a downwards trend for the average value of product_feature_1 over time, but this very slight.
- There does not appear to be any trend with product_feature_2.

Overall this means that it could be possible to see search date as an influencing factor for product_feature_1, product_feature_3 and product_feature_4. We would not expect to see if being an influencing factor for product_feature_1.

3.3 Data Transformation

3.3.1 One Hot Encoding

In order to apply any machine learning models to the data, we must first deal with the categorical variables. In the dataset there are only 2 distinctly categorical variables which are search_feature_4 and search_feature_5.

It could be argued that there are other categorical variables such as search_feature_1 and product_feature_2 which both have values ranging from 0 to 5, search_feature_3 which has values ranging from 0 to 340 with a step of 10 between each value, and even product_feature_1/3/4 which have continuous values from 0 to 67/38/39 respectively. Any of these fields could be considered as providing categorical information, but without any access to the business users we must assume that these are either ordinal fields (e.g. the star rating of a hotel where 5 stars is higher than 4 stars) or that these are numeric fields (e.g. the number of rooms required in an accommodation).

One method that can be used to convert the categorical features into binary values is known as one hot encoding. This effectively creates a binary column for each category within a field. This was applied to the search_feature_4 and search_feature_5 fields to create the following binary columns:

search_feature_4.A	search_feature_5.A
search_feature_4.B	search_feature_5.B
	search_feature_5.None

3.3.2 Feature Scaling

Scaling can be an important step as it ensures that features with smaller magnitudes do not dominate any models that are generated. For example, a feature that has large magnitude of values such as search_feature_2 may result in an apparently small coefficient in a linear regression model when predicting the outcome of a label with small values such as product_feature_5. This may lead to a potentially false interpretation that search_feature_2 is unimportant as a predictor for product_feature_5.

However, for this case study we will not look at the size of the coefficients alone, but rather their significance or importance to the model. For example, the t-statistic associated with each coefficient in the linear regression model will give a measure of feature importance. Therefore, no scaling of features is required.

3.4 Modelling

The goal of the case study is to find patterns and predictors of consumer behaviour. For the given dataset, consumer behaviour is effectively modelled by the characteristics of the products selected which is represented in the product_feature_1/2/3/4 fields. In order to find the predictors, we can fit various models for each of the product_feature_1/2/3/4 fields and determine which features have the highest significance or importance to the models.

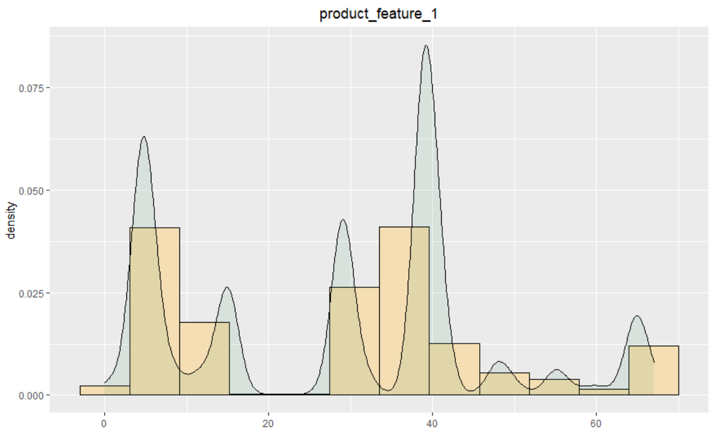
For example if we are looking at what drives the consumer selections for product_feature_1 we can fit a linear regression model using all of the available features and determine which of the features have the greatest impacts. We can then repeat the process for a decision tree model and a random forest model and compare the results. If the same features are reported as being important across all models then we can confidently state that these must be predictors for product_feature_1.

4 Results

4.1 Predictors for product_feature_1

4.1.1 Distribution

In order to better understand any results regarding product_feature_1, we can take a look at the distribution of its values using a histogram overlaid with a probability density plot.



The distribution appears to be multi-modal – i.e. there are multiple distinct peaks in the histogram and density plot.

Range: 0 to 67

Mean: 28.7

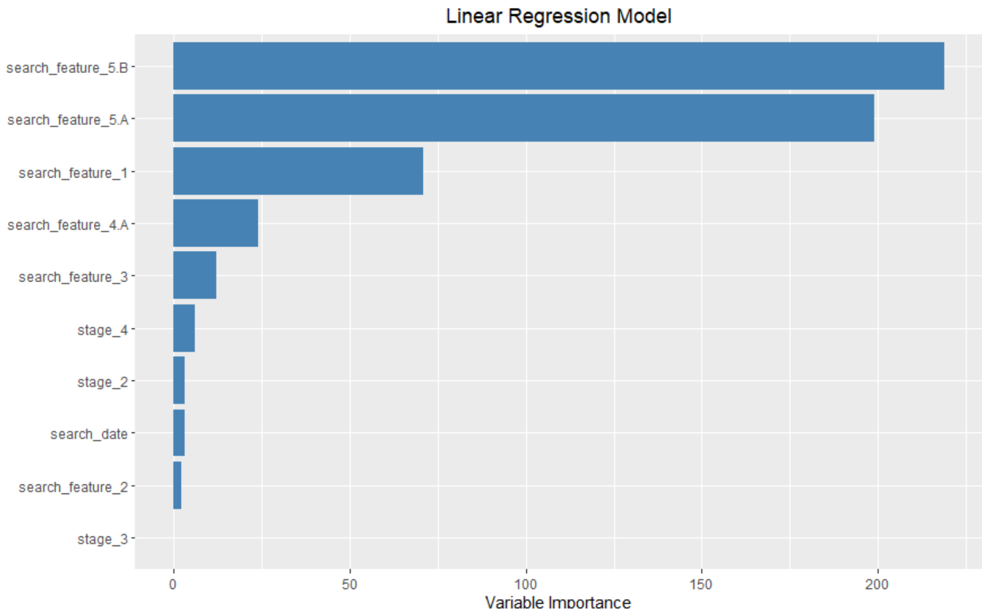
Standard Deviation: 18.6

4.1.2 Linear Regression Model

A linear regression model was fitted for product_feature_1 using the “lm” function in R and produced the following output:

```
Coefficients: (2 not defined because of singularities)
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         10.4027361  0.1548016  67.200  < 2e-16 ***
search_date         -0.0072547  0.0024823  -2.923  0.00347 **
stage_2              0.3708402  0.1191452   3.113  0.00186 **
stage_3             -0.0279727  0.0745638  -0.375  0.70755
stage_4             -1.3376327  0.2093160  -6.390 1.66e-10 ***
search_feature_1    -1.3490661  0.0191341 -70.506  < 2e-16 ***
search_feature_2     0.0007868  0.0003763   2.091  0.03654 *
search_feature_3     0.0083831  0.0006824  12.284  < 2e-16 ***
search_feature_4.A   1.4109187  0.0598347  23.580  < 2e-16 ***
search_feature_4.C          NA         NA      NA       NA
search_feature_5.A  22.2084110  0.1115297 199.126  < 2e-16 ***
search_feature_5.B  23.9129299  0.1090416 219.301  < 2e-16 ***
search_feature_5.None       NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.55 on 303782 degrees of freedom
Multiple R-squared:  0.2999,    Adjusted R-squared:  0.2998
F-statistic: 1.301e+04 on 10 and 303782 DF,  p-value: < 2.2e-16
```

We can see the list of coefficients, their standard errors, t statistics and p-values within the table. A view of the variable importance is shown in the chart below.



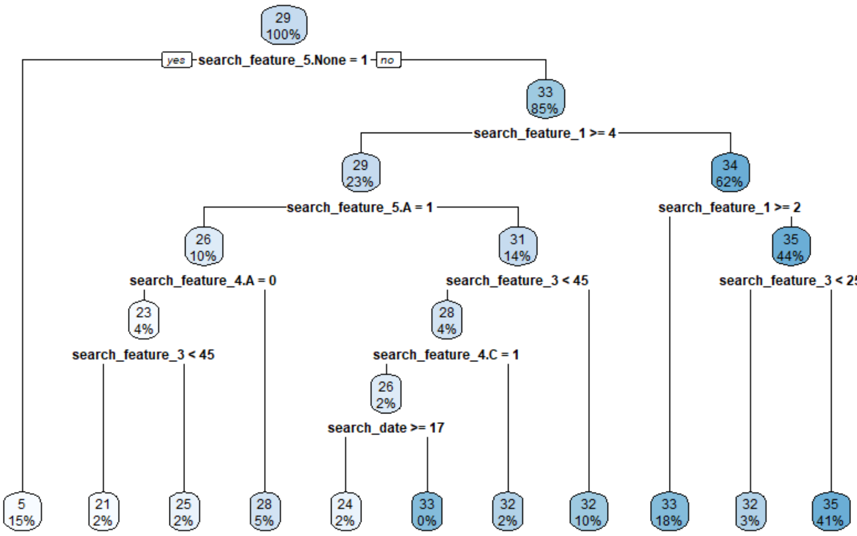
This shows that search_feature_5 is clearly the most important variable in the model, followed by search_feature_1, search_feature_4 and search_feature_3. In looking at the coefficients we can see that when a value of "A" or "B" is given for search_feature_5, the expected value of product_feature_1 increases by about 22 units. Similarly, when a value of "A" is given for search_feature_4, the expected value of product_feature_1 increases by about 1.4 units.

We can also see that each additional 10 units attributed to search_feature_3 increases the expected value of product_feature_1 by about 0.08 but each additional unit attributed to search_feature_1 reduces the expected value of product_feature_1 by about 1.35.

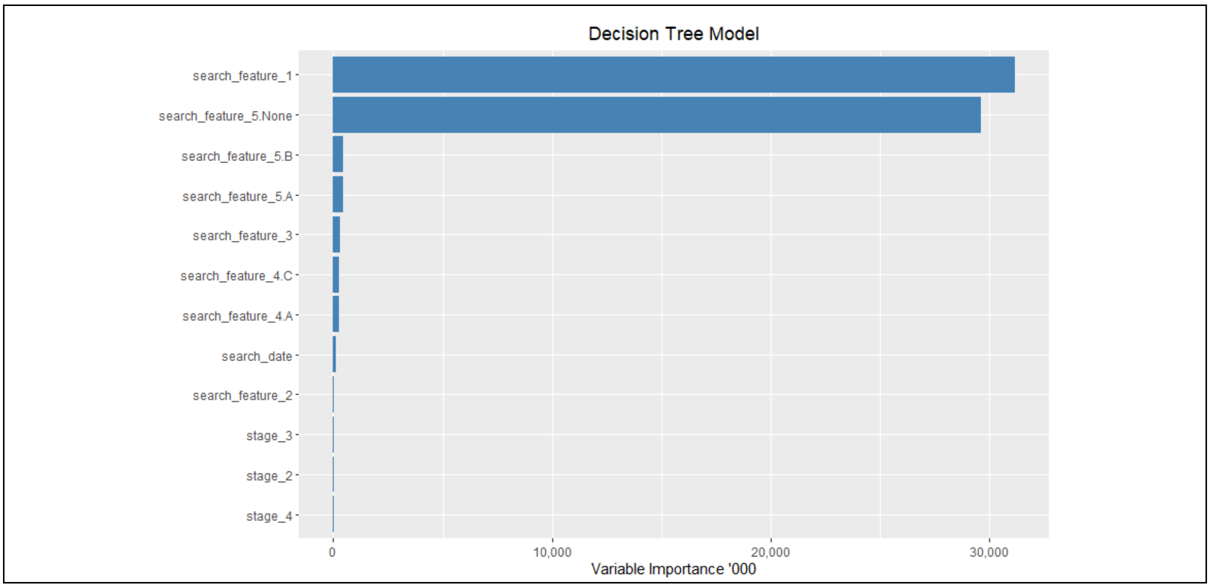
A measure of the degree by which the model can explain its variation is the R^2 statistic. In general terms the higher the value the better the model, because the model is able to explain more of the variation with its data. We can see from the output above that the value of R^2 in this case is **0.2999** which means that we only have a moderately good fitting model.

4.1.3 Decision Tree Regression Model

A different type of regression model that we can try is the decision tree model. A similar approach was undertaken, this time using the "rpart" function in R and the resulting tree is shown below.



This decision tree resulted in a R^2 score of **0.3073**, a slightly better result than we achieved from the linear regression model. From the tree we can see that the most important features are search_feature_5, followed by search_feature_1, search_feature_3 and search_feature_4. We can also produce a variable importance plot to show the magnitude of the importance in the generated model.

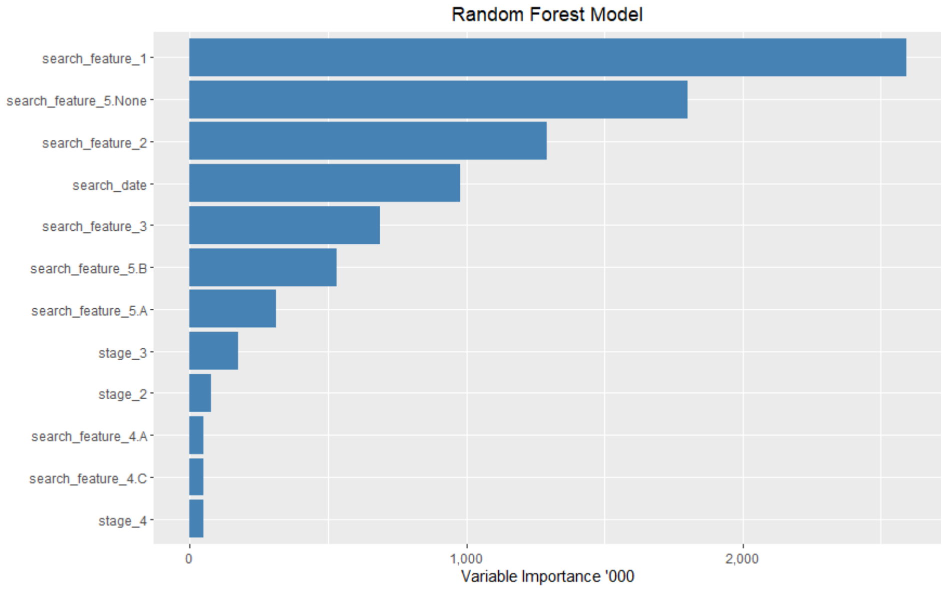


This quite clearly shows that the 2 most important features by far are search_feature_1 and search_feature_5, with some importance given to search_feature_3 and search_feature_4. This shows a similar tale to the results gleaned from the linear regression model.

4.1.4 Random Forest Regression Model

The third model to attempt is the random forest regression model. This takes the concept of decision trees and attempts to improve on the results through bootstrapping techniques. The "randomForest" function in R was used to model the data, although it must be noted that only a random sample of 50,000 records in the dataset was used due to limitations in the computing power available.

The random forest model resulted in R^2 score of **0.3322** which is an improvement on the decision tree model. However, one of the main drawbacks with a random forest is that it is a "black box" model and therefore difficult to interpret. We can still draw some conclusions through a plot of the variable importance.



In this case we could say that search_feature_1 has the most influence on the model, followed by search_feature_5, search_feature_2, search_date and search_feature_3. Interestingly, search_feature_4 is not considered to be of importance to the random forest model.

4.1.5 Conclusion – product_feature_1

Conclusions are drawn from looking at all 3 models generated. The following table shows the predictors that were consistently prominent across the different models.

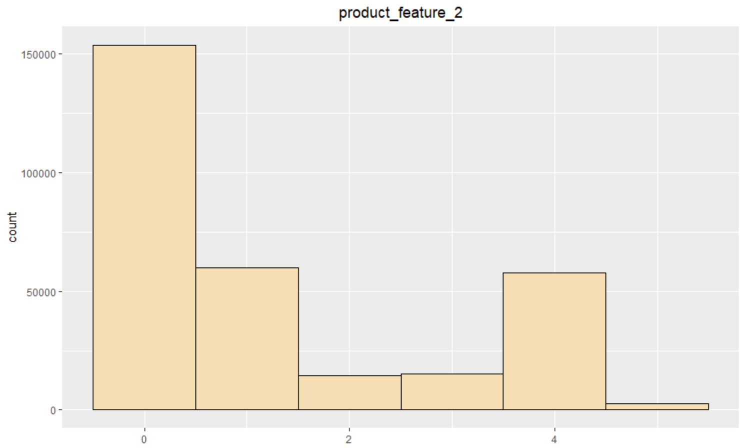
Predictor	Importance Rank	Relationship to product_feature_1*
search_feature_1	1	Inverse relationship: the greater the value for search_feature_1, the lower the value for product_feature_1.
search_feature_5	2	Positive relationship: a value of "A" or "B" will increase the value for product feature_1.
search_feature_3	3	Positive relationship: each increment of 10 units for search_feature_1 will increase the value for product_feature_1.

** when all other predictors are kept constant*

4.2 Predictors for product_feature_2

4.2.1 Distribution

In order to better understand any results regarding product_feature_2, we can take a look at the distribution of its values using a histogram.



The distribution shows that the majority of records have a value of 0, but not many records have values of 2, 3 or 5.

Range: 0 to 5

Mean: 1.2

Median: 0

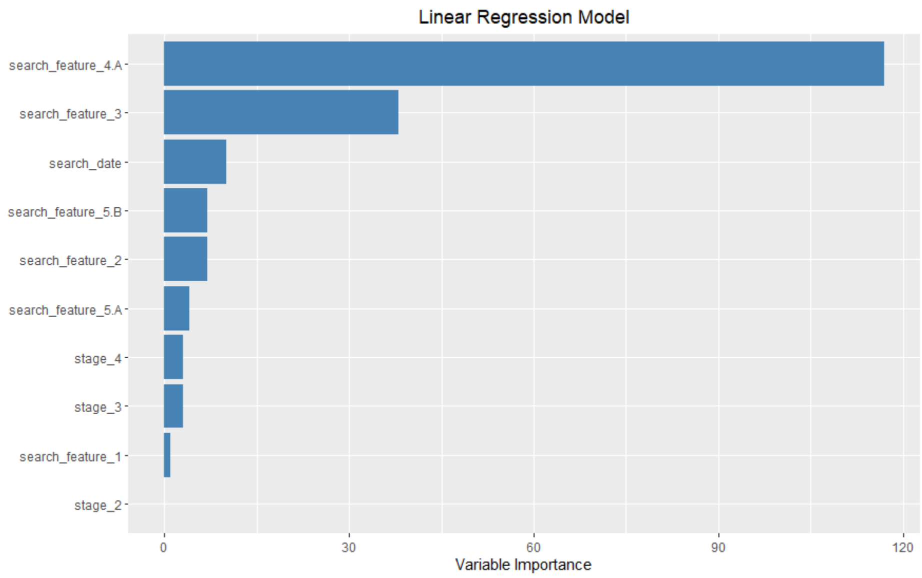
Standard Deviation: 1.59

4.2.2 Linear Regression Model

A linear regression model was fitted for product_feature_2 using the “lm” function in R and produced the following output:

```
Coefficients: (2 not defined because of singularities)
                      Estimate Std. Error  t value Pr(>|t|)    
(Intercept)          1.878e+00  1.546e-02  121.435  < 2e-16 ***
search_date          2.485e-03  2.479e-04   10.022  < 2e-16 ***
stage_2             -8.154e-04  1.190e-02   -0.069  0.94537    
stage_3             -2.023e-02  7.447e-03   -2.716  0.00661 ** 
stage_4              6.555e-02  2.091e-02    3.135  0.00172 ** 
search_feature_1     1.227e-03  1.911e-03    0.642  0.52083    
search_feature_2     2.529e-04  3.759e-05    6.730 1.70e-11 ***
search_feature_3    -2.586e-03  6.816e-05  -37.941  < 2e-16 ***
search_feature_4.A  -6.970e-01  5.976e-03 -116.626  < 2e-16 ***
search_feature_4.C          NA         NA       NA       NA    
search_feature_5.A  -4.502e-02  1.114e-02   -4.041 5.32e-05 ***
search_feature_5.B  -7.760e-02  1.089e-02   -7.125 1.04e-12 ***
search_feature_5.None       NA         NA       NA       NA    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.553 on 303782 degrees of freedom
Multiple R-squared:  0.05243,	Adjusted R-squared:  0.05239 
F-statistic:  1681 on 10 and 303782 DF,  p-value: < 2.2e-16
```

We can see the list of coefficients, their standard errors, t statistics and p-values within the table. The model produced a R^2 value of **0.0524** which means that overall the model only explains 5% of the variation in the data. This is a poor predictive model, but nevertheless a view of the variable importance is shown in the chart below.



This shows that search_feature_4 is clearly the most important variable in the model, followed by search_feature_3, search_date, search_feature_5 and search_feature_2. In looking at the coefficients we can see that when a value of "A" is given for search_feature_4, the expected value of product_feature_2 decreases by about 0.7 units.

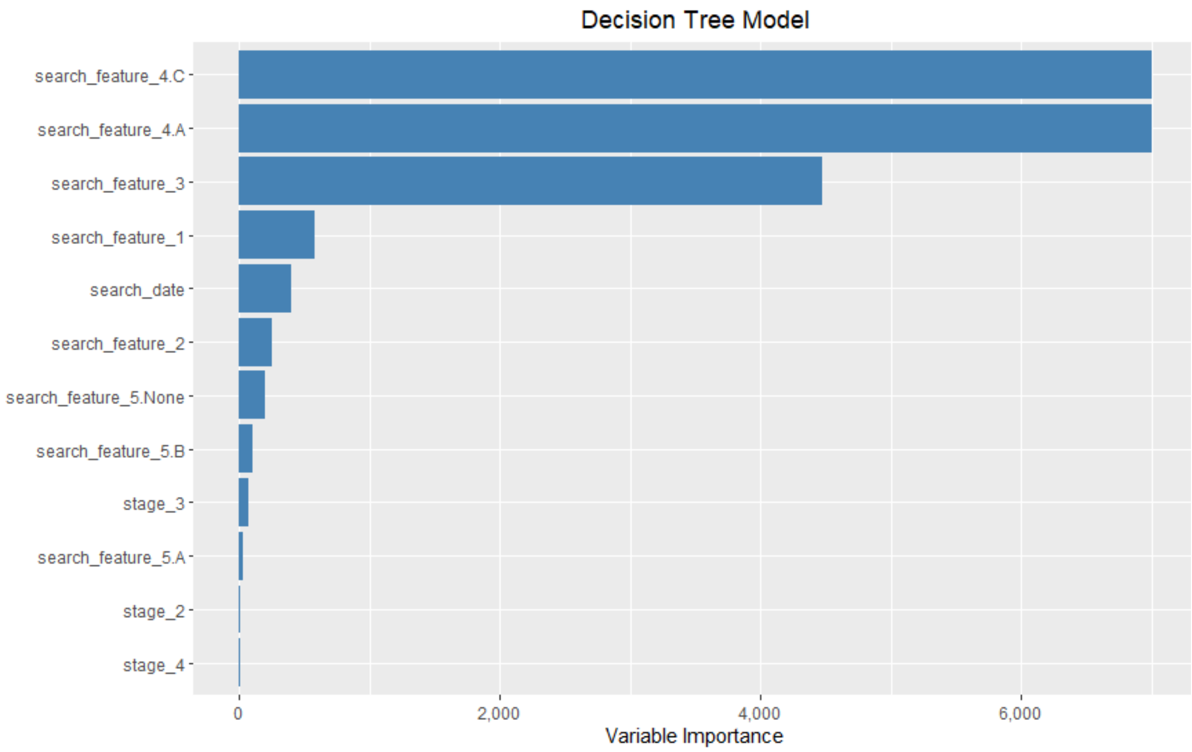
4.2.3 Decision Tree Classification Model

Following the poor result of the linear regression model a decision tree classification model was fitted. In this model the label is treated as categorical rather than numeric, with no ordinal assumptions made about the categories. The results are shown in the confusion matrix below.

```
Confusion Matrix and Statistics

          Reference
Prediction      0      1      2      3      4      5
         0 146134  47587  13461  14292  41253   2492
         1    859   2203    153    138   1607     28
         2    129    242    302    107    176      2
         3      0      0      0      0      0      0
         4   6654   9750    719    534  14876     95
         5      0      0      0      0      0      0

Overall Statistics

               Accuracy : 0.5382
                 95% CI : (0.5365, 0.54)
```

As this is now a classification model, there is no R^2 score. Instead we can get a score of Accuracy, which is the proportion of instances where the model correctly predicts the value for product_feature_2. Here we see that the Accuracy is 0.5382 – i.e. the correct result is predicted in just over half of the records. Interestingly we can see that the model made no predictions for values of either 3 or 5. Although this is not a great model it is an improvement on the linear regression results.

We can produce a variable importance plot to show the magnitude of the importance in the generated model.



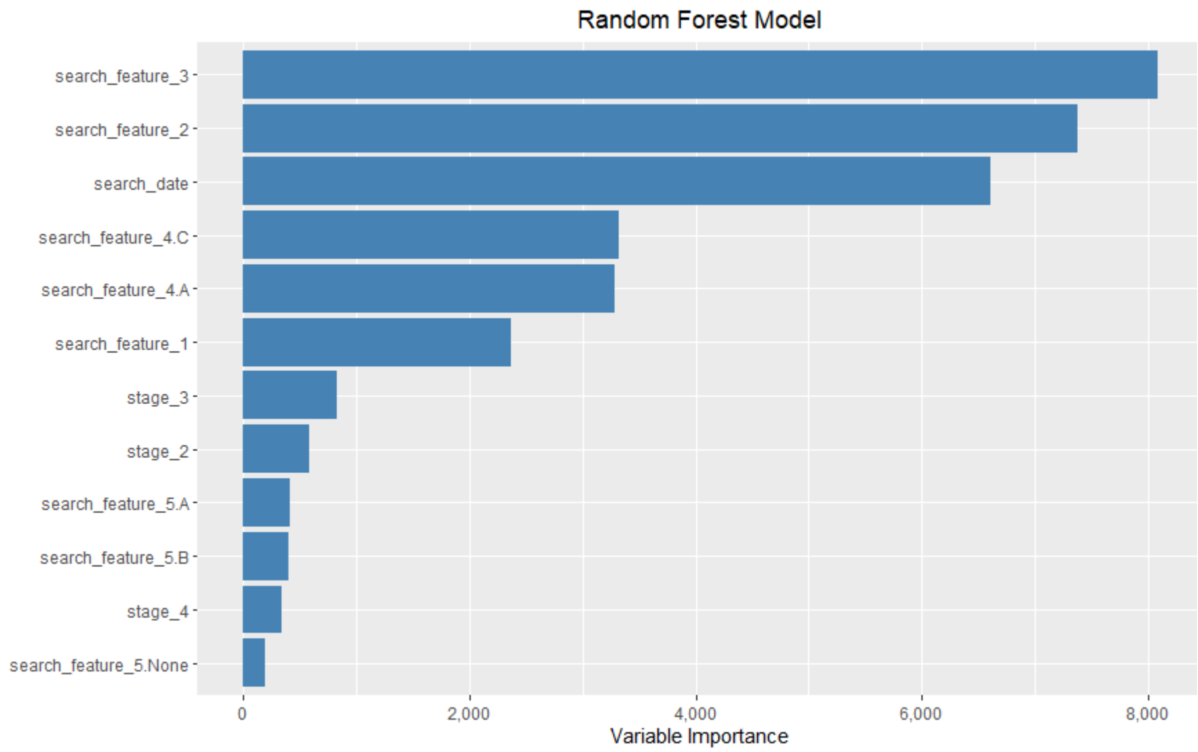
This quite clearly shows that the 2 most important features by far are search_feature_4 and search_feature_3 with some importance given to search_feature_1 and search_date. This shows a similar tale to the results gleaned from the linear regression model.

4.2.4 Random Forest Classification Model

Following the questionable results from the linear regression and decision tree classification model a random forest classification model was fitted. The results are shown in the confusion matrix below.

```
Confusion Matrix and Statistics

          Reference
Prediction      0      1      2      3      4      5
         0 147470  45167  13074  14127  39402   2485
         1   1007   6978    218    244   1845     62
         2     22     47    886     29     78      0
         3      1      4      0    230      1      1
         4   5276   7586    457    441  16586     68
         5      0      0      0      0      0      1

Overall Statistics

               Accuracy : 0.5667
                 95% CI : (0.5649, 0.5684)
```

The random forest model resulted in an Accuracy score of **0.5667** which is an improvement on the decision tree model, and we can see that it was able to correctly predict a value of 3 on 230 occasions compared to 0 for the decision tree model. A plot of the variable importance is below.



This model tells us that search_feature_3, search_feature_2 and search_date seem to have the most influence, while search_feature_4 and search_feature_1 also have some influence on the model.

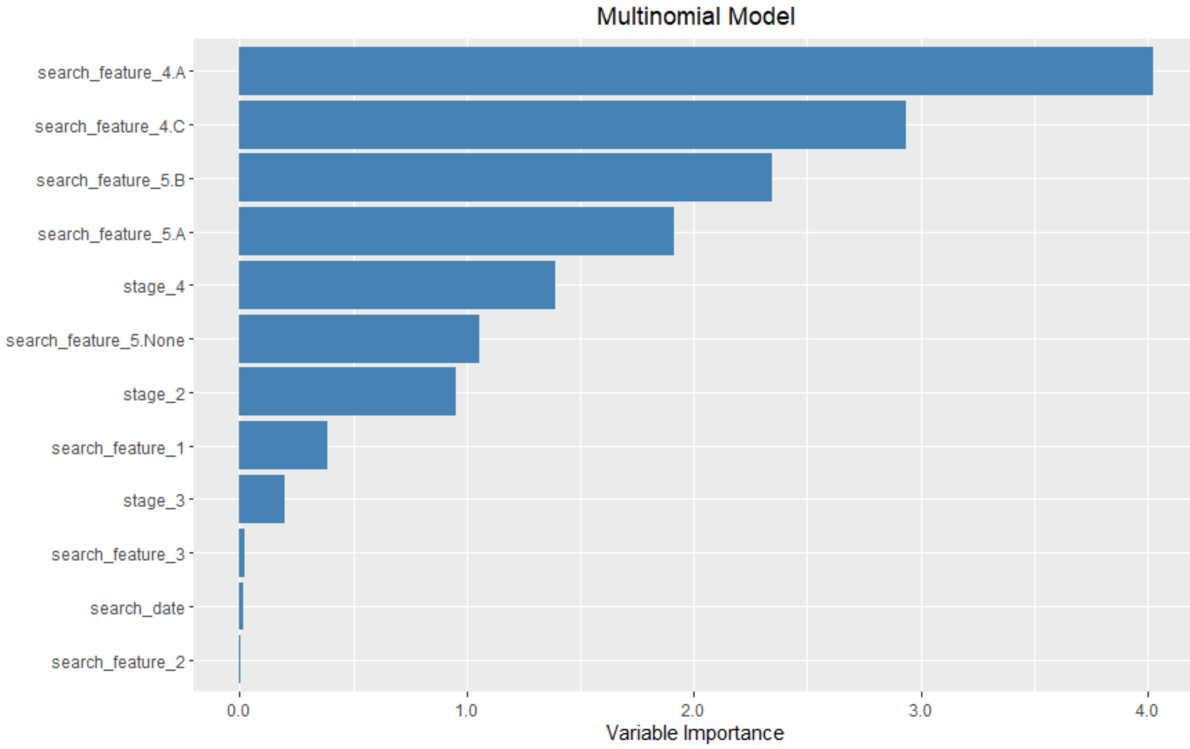
4.2.5 Multinomial Classification Model

A fourth model can be considered here, whereby the ordinality of product_feature_2 can be assumed. In other words, product_feature_2 is assumed to contain some sort of orderly ranking – e.g. a 4 star accommodation vs 3 star accommodation. This model is known as the Multinomial model.

The results of this model are shown in the confusion matrix below.

```
Confusion Matrix and Statistics

          Reference
Prediction      0      1      2      3      4      5
         0 142146  47383  13842  14235  40289   2505
         1    690   1473    101    123   1090     13
         2      0      3      3      2      1      0
         3      0      0      0      0      0      0
         4  10940  10923    689    711  16532     99
         5      0      0      0      0      0      0

Overall Statistics

               Accuracy : 0.5272
                 95% CI : (0.5254, 0.529)
```

The multinomial model resulted in an Accuracy score of **0.5272** which is a lower value than that both the decision tree and the random forest models. Interestingly no predictions were made for a value of either 3 or 5. A plot of the variable importance is below.



This model tells us that search_feature_4, search_feature_5 and stage_4 seem to have the most influence, while stage_2 and search_feature_1 also have some influence on the model.

4.2.6 Conclusion – product_feature_2

Conclusions are drawn from looking at all 4 models generated. The following table shows the predictors that were consistently prominent across the different models.

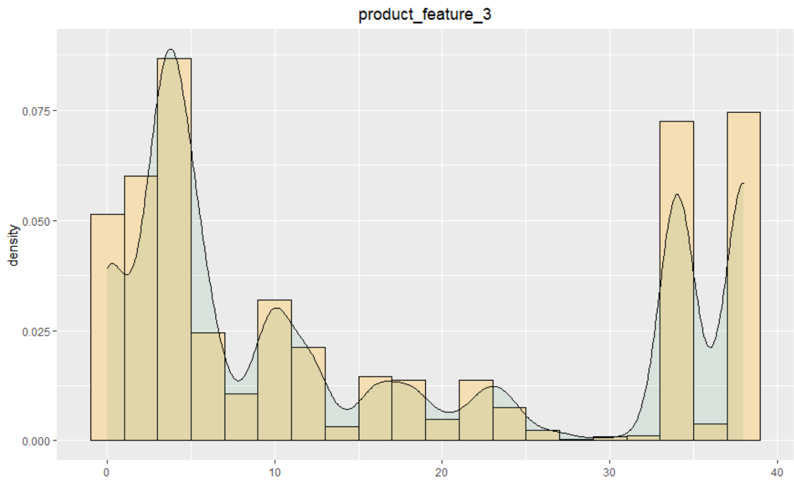
Predictor	Importance Rank	Relationship to product_feature_2*
search_feature_4	1	Inverse relationship: a value of “A” or “C” will decrease the value for product_feature_2.
search_feature_3	2	Inverse relationship: each increment of 10 will decrease the value for product_feature_2.
search_date	3	Inverse relationship: each increment of search_date will decrease the value for product_feature_2.
search_feature_2	4	Inverse relationship: each increment of search_feature_2 will decrease the value for product_feature_2.

** when all other predictors are kept constant, garnered from linear regression model where intercept (starting value for product_feature_2) = 1.88*

4.3 Predictors for product_feature_3

4.3.1 Distribution

In order to better understand any results regarding product_feature_3, we can take a look at the distribution of its values using a histogram overlaid with a density plot.



The histogram and density plot shows a multi-modal distribution with values skewed at both ends of the range.

Range: 0 to 38

Mean: 16.2

Median: 10

Standard Deviation: 14.35

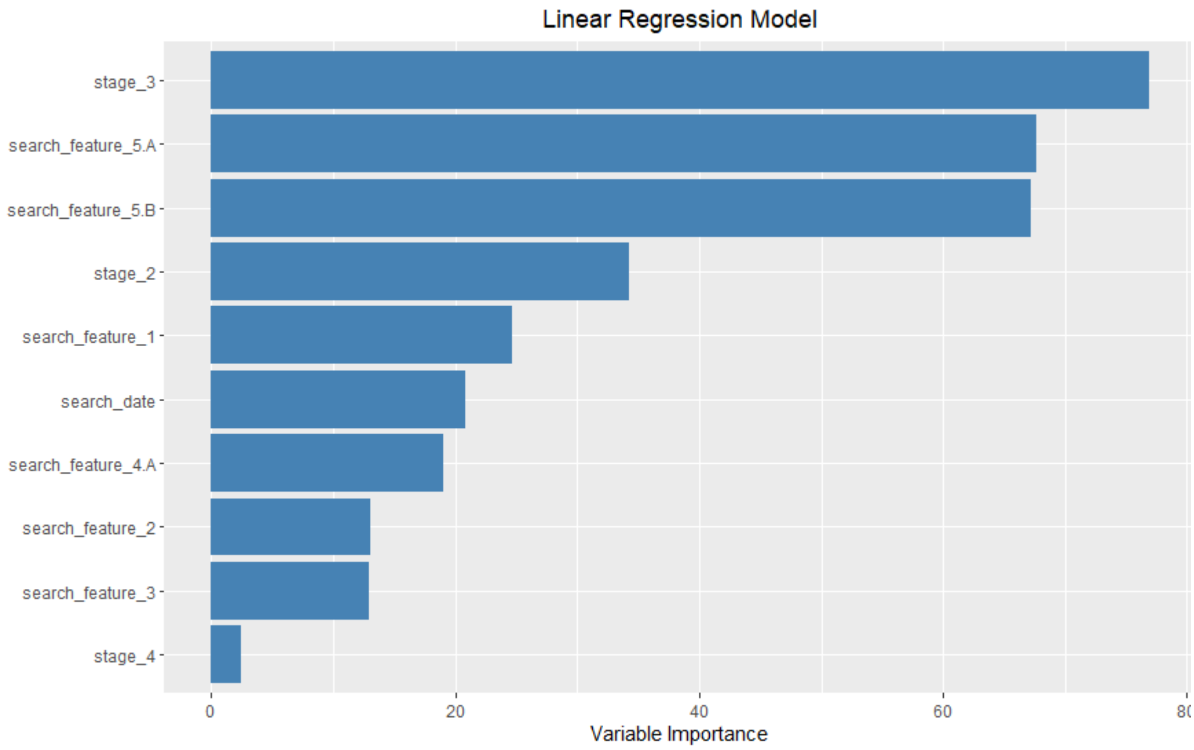
The large value for standard deviation confirms the skewness of the data at each end of the range.

4.3.2 Linear Regression Model

A linear regression model was fitted for product_feature_3 using the "lm" function in R and produced the following output:

```
Call:
lm(formula = product_feature_3 ~ ., data = lh_work)

Residuals:
    Min      1Q  Median      3Q     Max 
-28.960 -10.988  -4.752  11.610  30.255 

Coefficients: (2 not defined because of singularities)
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)        25.3353489  0.1366029 185.467   <2e-16 ***
search_date         0.0456606  0.0021905  20.845   <2e-16 ***
stage_2            -3.5949664  0.1051383 -34.193   <2e-16 ***
stage_3            -5.0606629  0.0657980 -76.912   <2e-16 ***
stage_4            -0.4537301  0.1847085  -2.456    0.014 *  
search_feature_1    0.4163651  0.0168847  24.659   <2e-16 ***
search_feature_2   -0.0043446  0.0003321 -13.083   <2e-16 ***
search_feature_3   -0.0077813  0.0006022 -12.921   <2e-16 ***
search_feature_4.A -1.0021924  0.0528004 -18.981   <2e-16 ***
search_feature_4.C         NA         NA      NA       NA    
search_feature_5.A -6.6580027  0.0984181 -67.650   <2e-16 ***
search_feature_5.B -6.4671288  0.0962225 -67.210   <2e-16 ***
search_feature_5.None      NA         NA      NA       NA    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.72 on 303782 degrees of freedom
Multiple R-squared:  0.08586,	Adjusted R-squared:  0.08583 
F-statistic:  2853 on 10 and 303782 DF,  p-value: < 2.2e-16
```

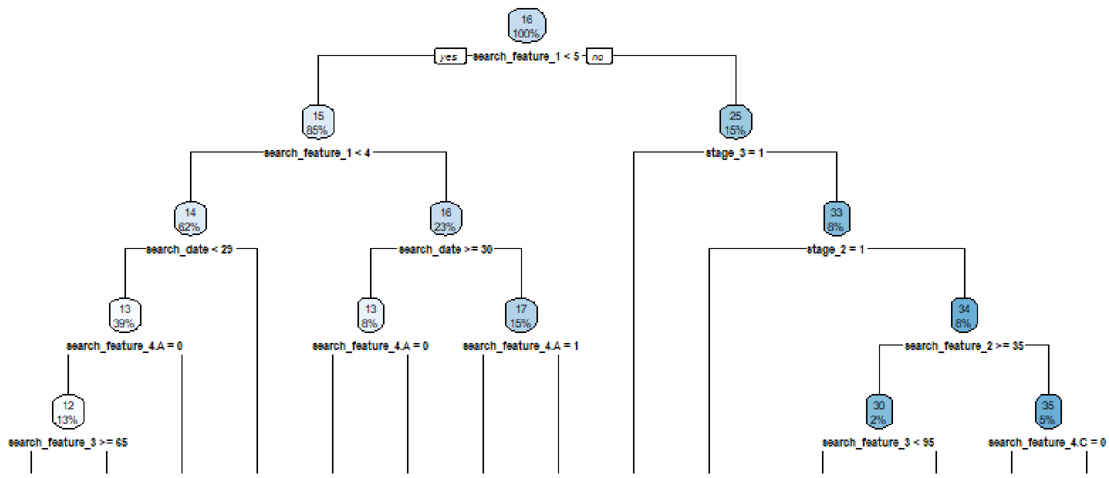
We can see the list of coefficients, their standard errors, t statistics and p-values within the table. A view of the variable importance is shown in the chart below. We can see from the output above that the value of R^2 is **0.0859** which means that we have a poor model and therefore the variable importance given is questionable.



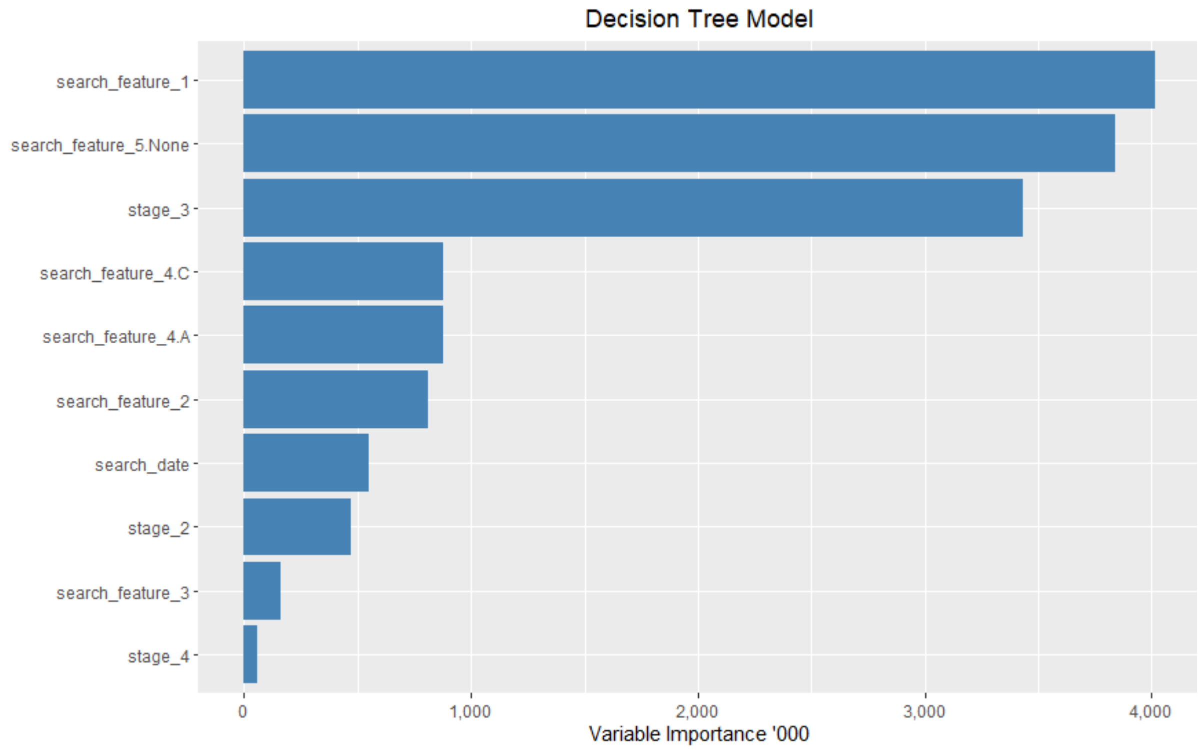
This shows that stage_3 and search_feature_5 are clearly the most important variables in the model, followed by stage_2, search_feature_1, search_date and search_feature_4. In looking at the coefficients we can see that when the user gets to stage_3 of the process, the expected value of product_feature_3 decreases by about 5 units. Similarly, when a value of “A” or “B” is given for search_feature_5, the expected value of product_feature_3 decreases by about 6.5 units.

4.3.3 Decision Tree Regression Model

Following the poor result of the linear regression model a decision tree regression model was fitted. A snapshot of the top section of the resulting tree is shown below:



This decision tree resulted in a R^2 score of **0.1457**, a slightly better result than we achieved from the linear regression model but still indicates that the model is not a great fit. From the tree we can see that the important features at the top of the tree are search_feature_1, stage_3, search_date and stage_2. Interestingly search_feature_5 does not seem to be of importance in this decision tree. However we can also produce a variable importance plot to show the magnitude of the importance in the generated model which is a more reliable indicator.

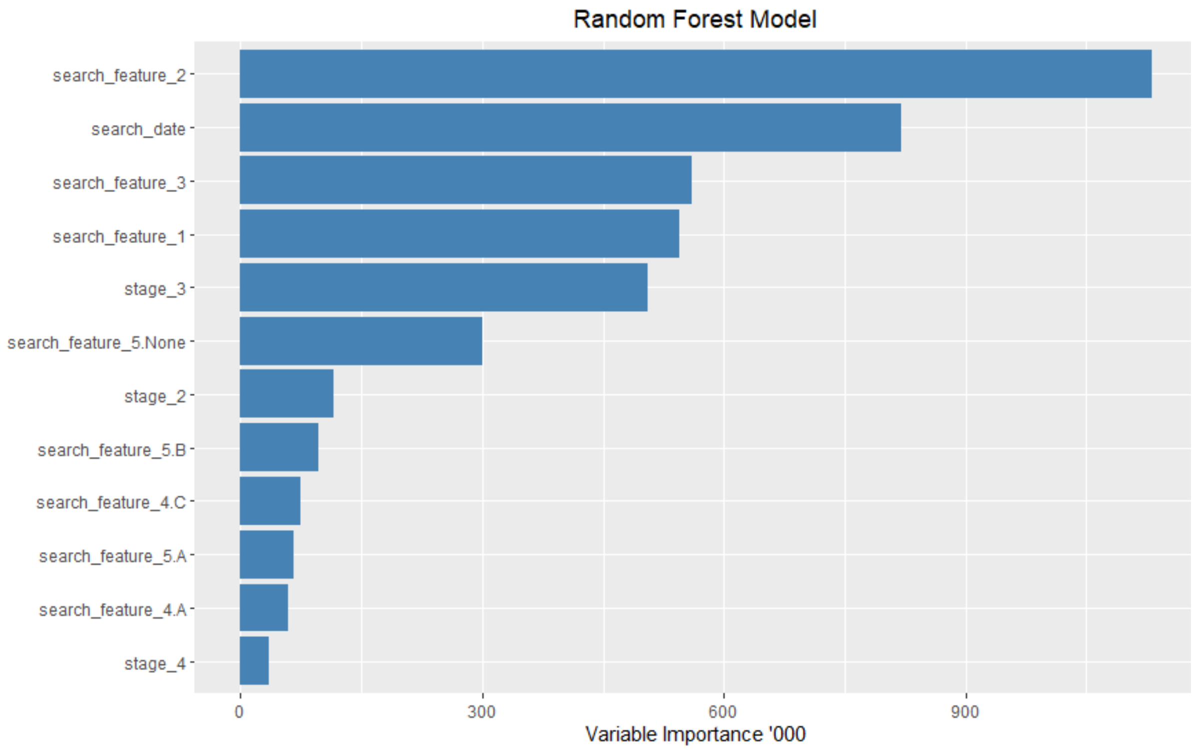


This quite clearly shows that the 3 most important features by far are search_feature_1, search_feature_5 and stage_3 with some importance given to search_feature_4 and search_feature_2. This shows similar results to those from the linear regression model.

4.3.4 Random Forest Regression Model

Following the results from the linear regression and decision tree regression model a random forest regression model was fitted. The resulting R^2 score was **0.1688** which is a slight improvement on the decision tree model. However, one of the main drawbacks with a random forest is that it is a "black box" model and therefore difficult to interpret. We can still draw some conclusions through a plot of the variable importance.

A plot of the variable importance is below.



This model tells us that search_feature_2 and search_date seem to have the most influence, while search_feature_3, search_feature_1 and stage_3 also have some influence on the model.

4.3.5 Naïve Bayes Model

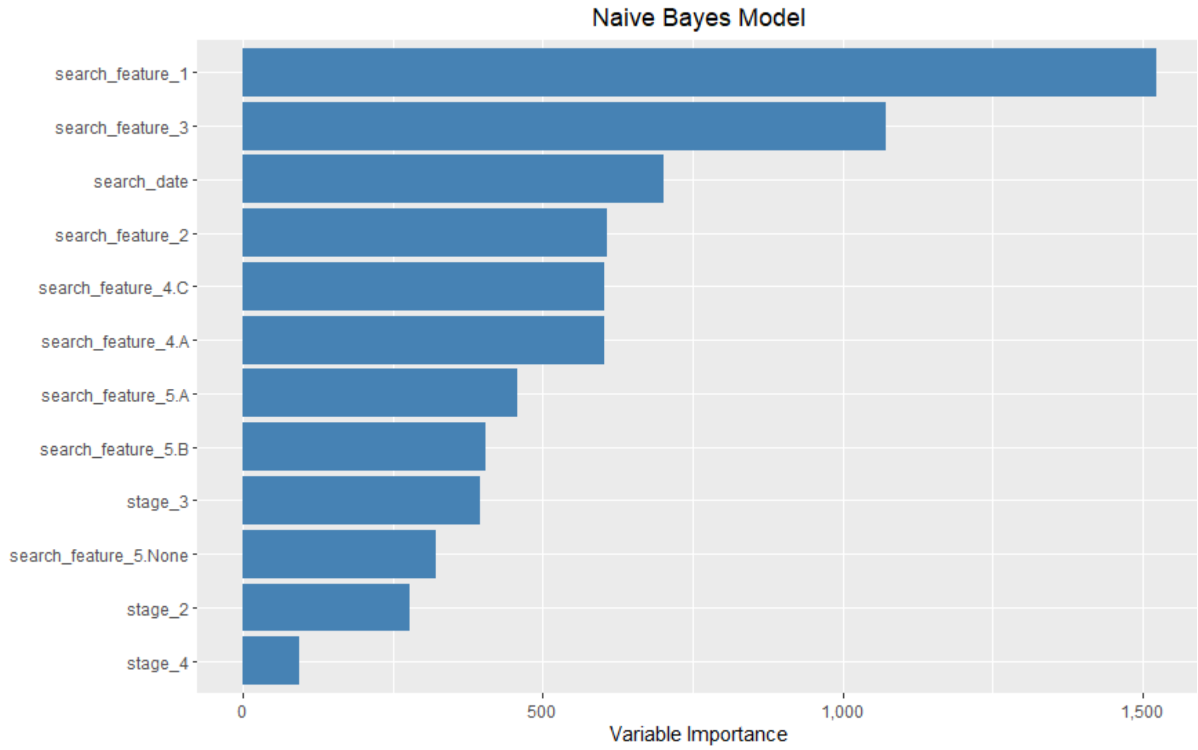
Due to the poor performance of the regression models for product_feature_3, an alternative approach was attempted. This involved changing the data to look at a classification model, rather than a numeric model, and then applying the Naïve Bayes algorithm to fit a model.

Naïve Bayes assumes that all features are independent and uses probabilistic techniques to build the model. The resulting model presented the following confusion matrix (first 10 values only):

```
Confusion Matrix and Statistics

          Reference
Prediction    0     1     2     3     4     5     6     7     8     9    10
        0    33     8     4    63    60    14    29     1     2     4    13
        1     0     0     0     0     0     0     0     0     0     0     0
        2     0     0     0     0     0     0     0     0     0     0     0
        3   452   136   212   843   652   281   236    74    27   141   256
        4  7846  1477  1485  9285 13376  5600  4346   944   478  1764  4453
        5     0     0     0     0     0     0     0     0     0     0     0
        6     0     0     0     0     0     0     0     0     0     0     0
        7   676   122   115   833  1013   482   403   106    42   133   345
        8     0     0     0     0     0     0     0     0     0     0     0
        9     0     0     0     0     0     0     0     0     0     0     0
        10 1049   181   125  1067   969   456   435    34    48   146  1124
```

This does not appear to be particularly accurate and the overall Accuracy score of **0.1934** is a reflection of this. However despite the low score, it is still likely to be a better prediction model than any of the linear regression, decision tree and random forest models.

A view of the variable importance is provided below:



This model tells us that search_feature_1 and search_feature_3 seem to have the most influence, while search_date, search_feature_2 and search_feature_4 also have some influence on the model.

4.3.6 Conclusion – product_feature_3

Conclusions are drawn from looking at all 3 models generated. The following table shows the predictors that were consistently prominent across the different models.

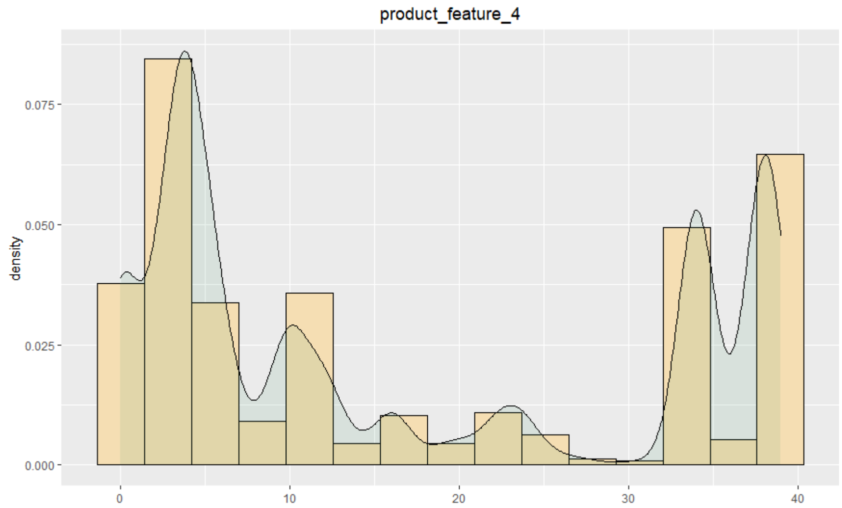
Predictor	Importance Rank	Relationship to product_feature_3*
search_feature_1	1	Positive relationship: each increment of 1 will increase the value of product_feature_3
stage_3	2	Inverse relationship: a value of 1 will decrease the value for product_feature_3.
search_feature_2	3	Inverse relationship: each increment of search_feature_2 will decrease the value for product_feature_3

** when all other predictors are kept constant, garnered from linear regression model where intercept (starting value for product_feature_3) = 25*

4.4 Predictors for product_feature_4

4.4.1 Distribution

In order to better understand any results regarding product_feature_4, we can take a look at the distribution of its values using a histogram overlaid with a density plot.



The histogram and density plot shows a multi-modal distribution with values skewed at both ends of the range.

Range: 0 to 39

Mean: 16.9

Median: 10

Standard Deviation: 14.85

The large value for standard deviation confirms the skewness of the data at each end of the range.

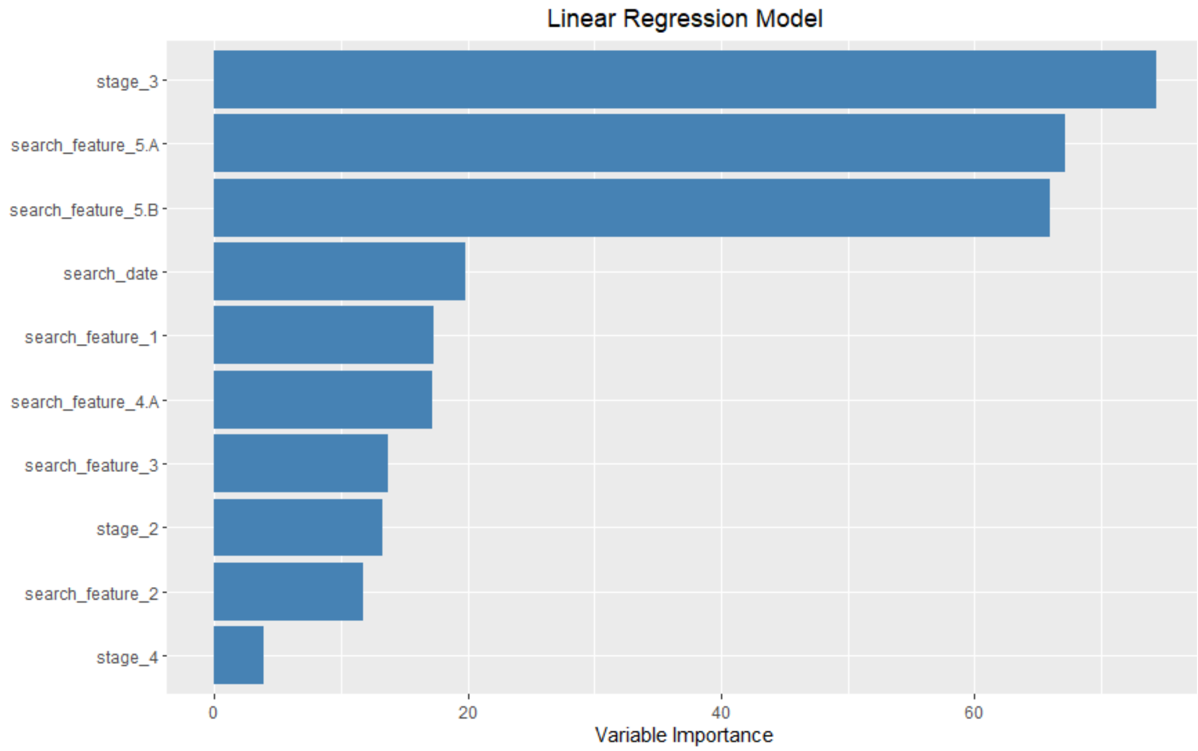
As suspected from the earlier analysis on correlation, this histogram and density plot, and indeed the summary statistics are very similar to the histogram for product_feature_3. This indicates that the models may show similar behaviour with regards to the feature/variable importance.

4.4.2 Linear Regression Model

A linear regression model was fitted for product_feature_4 using the "lm" function in R and produced the following output:

```
lm(formula = product_feature_4 ~ ., data = lh_work)

Residuals:
   Min     1Q Median     3Q    Max 
-29.35 -11.65  -5.13  14.22  28.74 

Coefficients: (2 not defined because of singularities)
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)           26.3128429  0.1420431 185.245  < 2e-16 ***
search_date            0.0450107  0.0022777  19.761  < 2e-16 ***
stage_2               -1.4552902  0.1093255 -13.312  < 2e-16 ***
stage_3               -5.0879578  0.0684184 -74.365  < 2e-16 ***
stage_4               -0.7379185  0.1920645  -3.842 0.000122 ***
search_feature_1       0.3042941  0.0175571  17.332  < 2e-16 ***
search_feature_2      -0.0040529  0.0003453 -11.737  < 2e-16 ***
search_feature_3      -0.0085795  0.0006262 -13.701  < 2e-16 ***
search_feature_4.A    -0.9428898  0.0549032 -17.174  < 2e-16 ***
search_feature_4.C            NA         NA      NA       NA    
search_feature_5.A    -6.8734406  0.1023376 -67.164  < 2e-16 ***
search_feature_5.B    -6.6014016  0.1000546 -65.978  < 2e-16 ***
search_feature_5.None         NA         NA      NA       NA    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.27 on 303782 degrees of freedom
Multiple R-squared:  0.07673,	Adjusted R-squared:  0.0767 
F-statistic:  2525 on 10 and 303782 DF,  p-value: < 2.2e-16
```

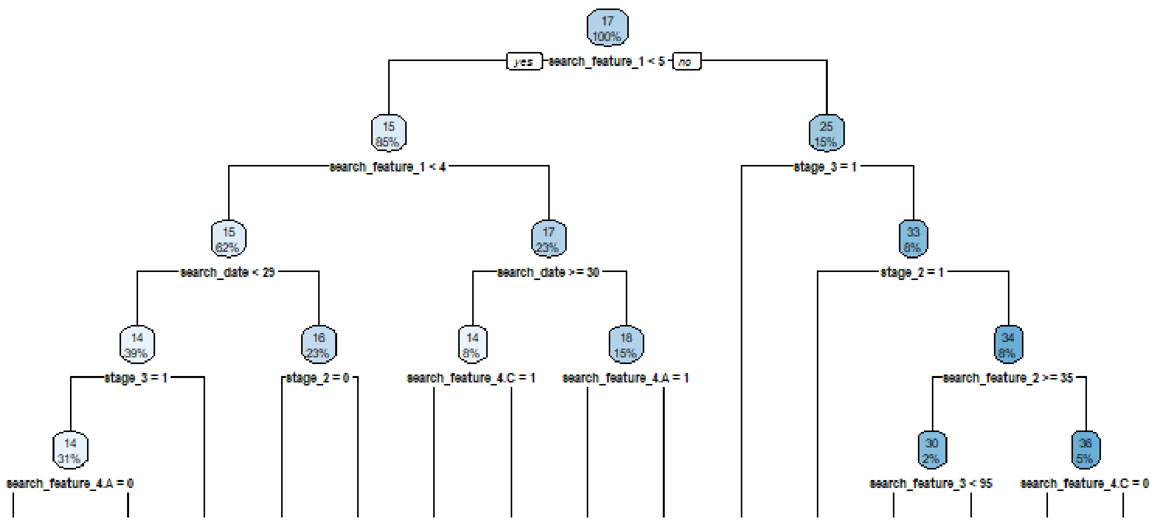
We can see the list of coefficients, their standard errors, t statistics and p-values within the table. A view of the variable importance is shown in the chart below. We can see from the output above that the value of R^2 is **0.0767** which means that we have a poor model and therefore the variable importance given is questionable.



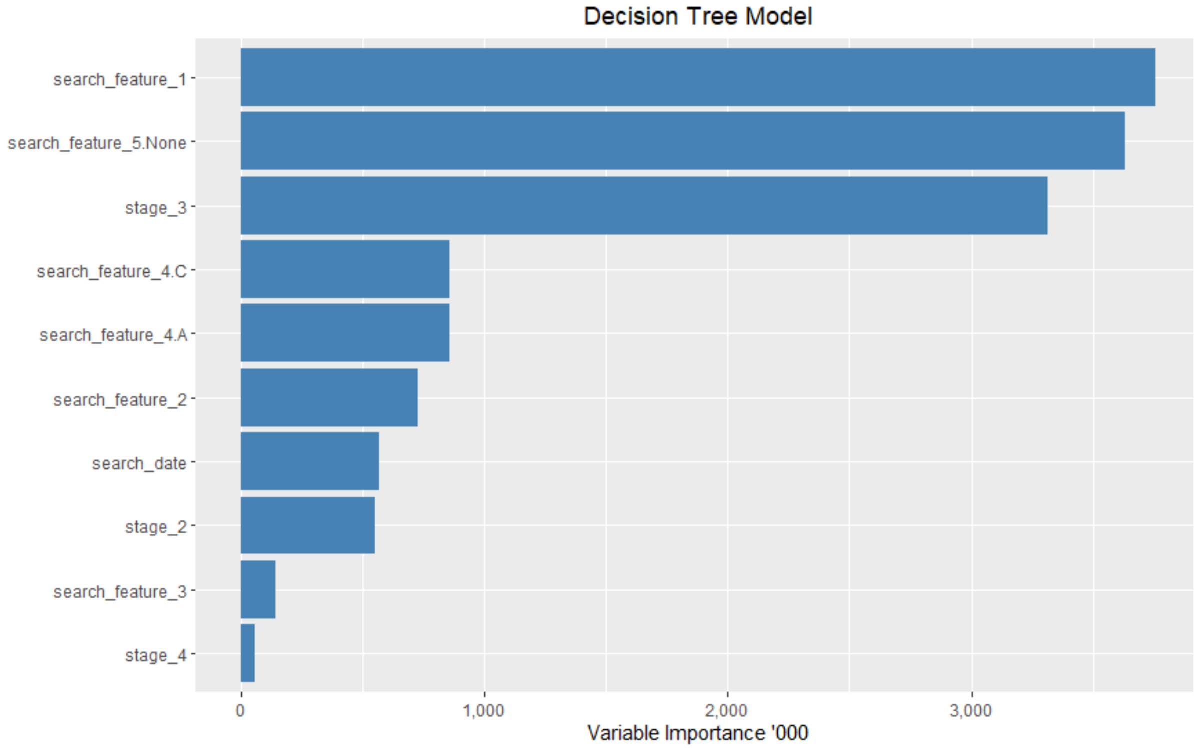
This shows that stage_3 and search_feature_5 are clearly the most important variables in the model, followed by search_date, search_feature_1, and search_feature_4. In looking at the coefficients we can see that when the user gets to stage_3 of the process, the expected value of product_feature_4 decreases by about 5 units which is a similar result to that given for product_feature_3. Similarly, when a value of "A" or "B" is given for search_feature_5, the expected value of product_feature_4 decreases by about 6.7 units. Again this mirrors the linear regression model generated for product_feature_3.

4.4.3 Decision Tree Regression Model

Following the poor result of the linear regression model a decision tree regression model was fitted. A snapshot of the top section of the resulting tree is shown below:



This decision tree resulted in a R^2 score of **0.1308**, a slightly better result than we achieved from the linear regression model but still indicates that the model is not a great fit. From the tree we can see that the important features at the top of the tree are search_feature_1, stage_3, search_date and stage_2. Interestingly search_feature_5 does not seem to be of importance in this decision tree. However we can also produce a variable importance plot to show the magnitude of the importance in the generated model which is a more reliable indicator.

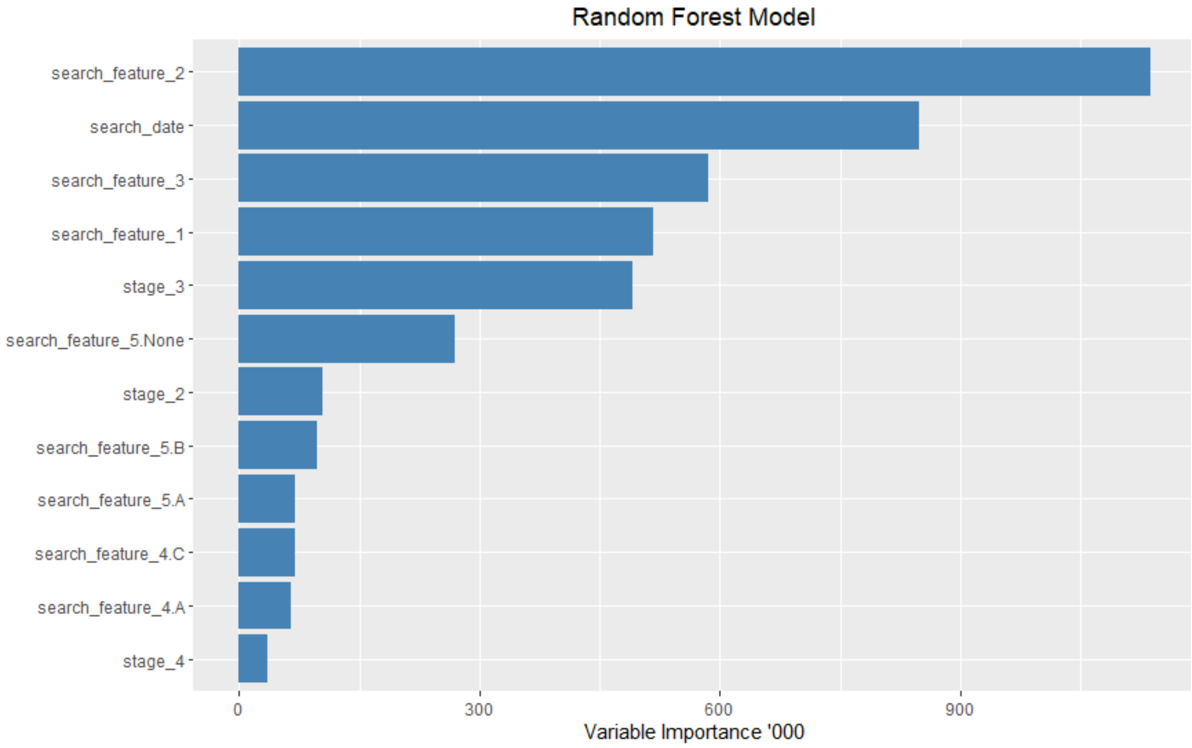


This quite clearly shows that the 3 most important features by far are search_feature_1, search_feature_5 and stage_3 with some importance given to search_feature_4 and search_feature_2. This shows similar results to those from the linear regression model and again is very similar to the variable importance from the decision tree model for product_feature_3.

4.4.4 Random Forest Regression Model

Following the results from the linear regression and decision tree regression model a random forest regression model was fitted. The resulting R^2 score was **0.1599** which is a slight improvement on the decision tree model. However, one of the main drawbacks with a random forest is that it is a "black box" model and therefore difficult to interpret. We can still draw some conclusions through a plot of the variable importance.

A plot of the variable importance is below.



This model tells us that search_feature_2 and search_date seem to have the most influence, while search_feature_3, search_feature_1 and stage_3 also have some influence on the model. Again, this is very similar to the random forest model generated for product_feature_3.

4.4.5 Naïve Bayes Model

Due to the poor performance of the regression models for product_feature_4, an alternative approach was attempted. This involved changing the data to look at a classification model, rather than a numeric model, and then applying the Naïve Bayes algorithm to fit a model.

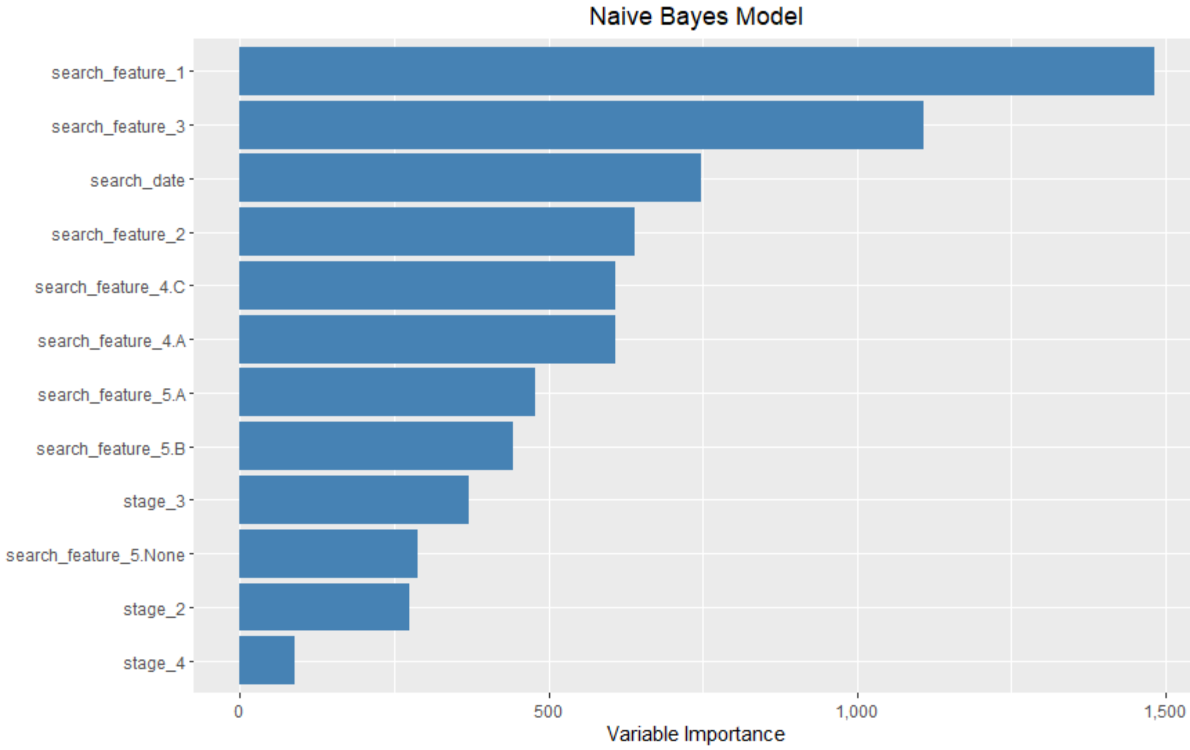
Naïve Bayes assumes that all features are independent and uses probabilistic techniques to build the model. The resulting model presented the following confusion matrix (first 10 values only):

Confusion Matrix and Statistics

Prediction \ Reference	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	19	20	21
0	0	0	0	0	2	0	0	0	0	0	1	0	1	0	0	0	1	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	12	2	3	13	2	7	6	0	1	2	3	0	2	2	1	0	2	0	0	0	1
4	1457	302	278	1678	2432	1034	798	142	90	234	778	135	498	109	83	21	464	48	97	102	49
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

This does not appear to be particularly accurate and the overall Accuracy score of **0.2094** is a reflection of this. However despite the low score, it is still likely to be a better prediction model than any of the linear regression, decision tree and random forest models.

A view of the variable importance is provided below:



This model tells us that search_feature_1 and search_feature_3 seem to have the most influence, while search_date, search_feature_2 and search_feature_4 also have some influence on the model.

4.4.6 Conclusion – product_feature_4

Conclusions are drawn from looking at 4 models generated. The following table shows the predictors that were consistently prominent across the different models.

Predictor	Importance Rank	Relationship to product_feature_4*
search_feature_1	1	Positive relationship: each increment of 1 will increase the value of product_feature_4
stage_3	=2	Inverse relationship: a value of 1 will decrease the value for product_feature_4
search_date	=2	Positive relationship: each increment of 1 will increase the value of product_feature_4
search_feature_5	4	Inverse relationship: a value of “A” or “B” for search_feature_5 will decrease the value for product_feature_4

** when all other predictors are kept constant, garnered from linear regression model where intercept (starting value for product_feature_4) = 26*