Section 1: Business Need and Importance

In today's fast-paced world, making healthy food choices can be challenging. With numerous factors influencing what constitutes a "healthy" option, individuals often find themselves overwhelmed by the complexity of nutritional information. While meal prepping offers a solution, the repetition of identical meals can deter many from committing to healthier eating habits. To address this issue, my analysis utilizes extensive datasets encompassing individual foods and complete recipes to develop a robust meal recommendation system. By leveraging supervised learning techniques and considering factors such as nutritional profiles, past eating behaviors, and potential allergies, this system aims to deliver tailored food recommendations and predict eating patterns effectively.

The goal is to simplify the process of selecting nutritious meals, empowering individuals to make informed choices effortlessly. By classifying foods based on user preferences and distinguishing between simple and complex recipes, the analysis aims to streamline meal planning and enhance overall dietary satisfaction. Additionally, the ability to differentiate between plant-based and animal-based foods aids in meeting the dietary restrictions of individual users. The insights derived from this analysis hold potential for meal preparation companies to refine their offerings and cater to the diverse nutritional needs of their clientele. Through this endeavor, I aspire to revolutionize the way we approach food consumption, promoting healthier lifestyles and facilitating greater accessibility to nutritious meals for all.

Section 2: Statistical Methodology

I employed multiple supervised learning techniques on historical data with predetermined outcomes. First, the Random Forest method was used to analyze a dataset containing 5000 recipes, with features including categorical variables representing ingredients such as beans, beef, eggs, peppers, onions, and sugar, as well as a binary label indicating the presence of dairy or nuts. This approach leveraged the ensemble of decision trees to identify significant predictors of recipe composition and allergen content, aiding in dietary recommendations for individuals with these specific dietary requirements. Using the K-Nearest Neighbor technique (KNN), I analyzed a dataset consisting of individual foods characterized by numerical nutritional profiles and binary indicators of consumption (eaten or rejected) based on survey results. By computing the similarity between food items based on their nutritional attributes, this method aids in developing personalized food recommendations tailored to an individual's dietary preferences and nutritional needs.

The Naive Bayes technique applied probabilistic reasoning to assess the likelihood of a recipe being either complex or simple, utilizing a dataset containing recipes categorized as one of these two options, with features including the presence or absence of specific ingredients encoded as binary variables. This approach provided insights into the factors contributing to recipe difficulty, guiding the classification of simplified recipes for novice cooks. Lastly, employing a Classification Tree model involved analyzing a dataset comprising over 5000 food items labeled as either vegan or non-vegan, with numerical features representing their nutritional composition. By constructing decision trees based on these features, this method enabled the prediction of food items' vegan status, supporting adherence to plant-based lifestyles.

Section 3: Results and Interpretation

Beginning with the Random Forest method, the goal was to develop a model for predicting if a recipe contains dairy or nuts. The model produced an accuracy of 57% which although better than a random selection, does not provide much confidence in its predictions. The Cumulative Lift Chart and Decile Wise Lift Chart (see Appendix) display this result in better detail. The Decile Wise Lift Chart, with a maximum gain value of 1, indicates that this model's minor improvement over random selection is not worth the cost of collecting more similar data. Figure 1 demonstrates the Variable Importance Plot, displaying which of the variables used in the analysis most influenced the resulting categorization. After reviewing this, I would consider testing the algorithm on different ingredients to see if they are better predictors of the presence of dairy or nuts.

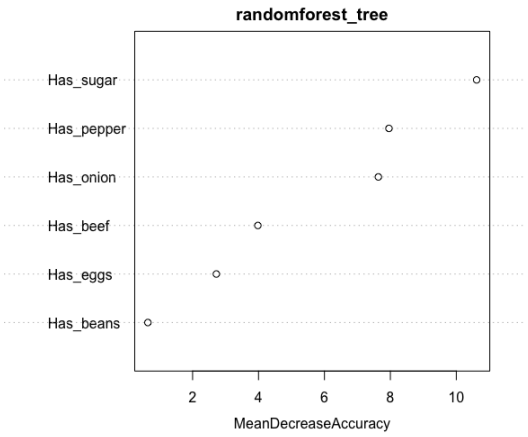


Figure 1: Variable importance plot for Random Forest

Next, I utilized the K-Nearest Neighbor method to analyze foods that were eaten by survey participants and foods that were not. This method allowed me to categorize these foods based on their nutritional content. The confusion matrix results (see Appendix) show that this method also does only slightly better than random selection, with an accuracy of 0.5088. The specificity (0.4994) of this model suggests that it is worse than random selection at correctly classifying meals that would

not be eaten by these individuals. These results suggest that nutritional content may not be a great indicator of eating patterns, and additional data may be needed to train an efficient model.

Naive Bayes was used to determine the classification of a recipe as simple or complex. Although these results again only proved to be slightly better at prediction than random, the specificity of this model (0.7094) suggests that my model does a better job of accurately predicting simple recipes than accurately predicting complex recipes (sensitivity of 0.5100). This can be seen in the Receiver Operator Curve for my model (see Appendix) which has an AUC of 0.6489. From these results, we can conclude that this model is useful in predictions about 65% of the time. In a low-risk situation such as this, the model is respectable enough to consider for use.

The Classification Tree method was able to build the most successful model of all my methods utilized. The best-pruned tree (Figure 2) gives a clear set of decision nodes to determine whether a food is Vegan (1) or not (0) based on its nutritional values. This model's resulting accuracy was 0.9713 and is supported by an AUC of 0.9812, both of which demonstrate this tree's effectiveness. In also looking at the decile-wise lift chart (see Appendix), we see that we only need about half of the data used to produce a model twice as efficient as random selection.

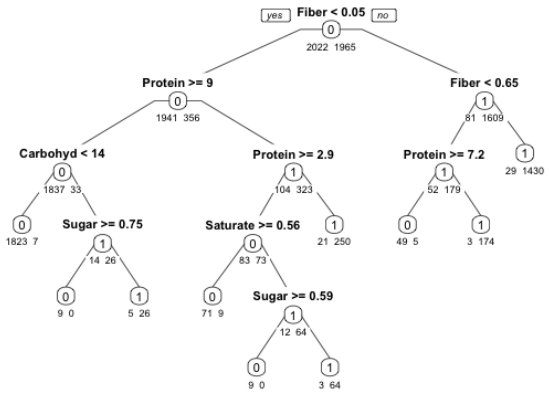


Figure 2: Best Pruned Classification Tree for predicting Vegan foods based on nutritional values

Section 4: Alternative Approaches

In determining alternative approaches for my statistical analysis, I considered applying Principal Component Analysis (PCA) to my data. The food dataset I used is a subset of a larger dataset that contains over 200 columns representing a vast assortment of nutritional values for each food. PCA would have allowed me to reduce the dimensionality of that full dataset, while still preserving most of the variance in the data, to then use linear regression in determining which nutrients have the most influence on calories. Although this method was a viable option, PCA's downfall is that in reducing dimensionality you lose information. In dietary analysis, it is essential to retain as much relevant information as possible to accurately characterize the nutritional properties of foods or recipes.

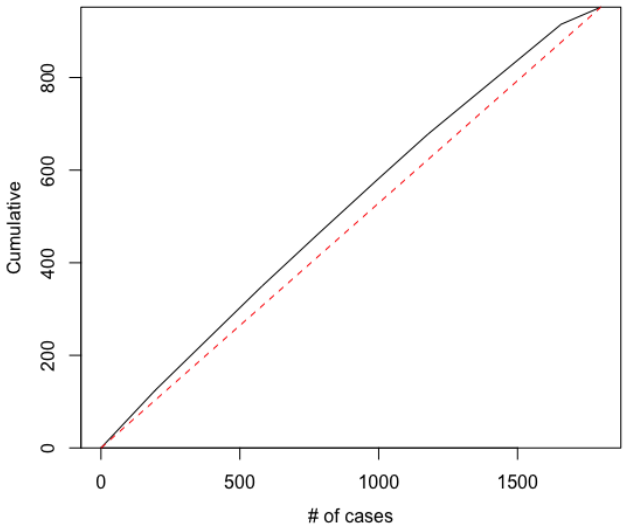
Section 5: Conclusions

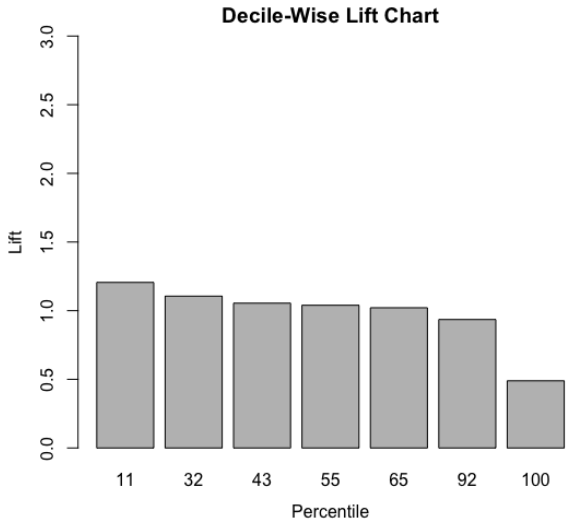
In today's market, demand for personalized dietary guidance is high, prompting the need for advanced data analytics. By analyzing vast food datasets, we can transform how individuals make dietary choices. Eating patterns are influenced by numerous factors beyond nutritional content, including health goals, dietary restrictions, geographic location, and religious affiliation that will determine whether a food is consumed. The results of my analysis prove how valuable the types of information collected can be to building valuable models. There will need to be much further analysis of foods, recipes, and the individuals whom the system is recommending foods for. With these insights, businesses can develop innovative meal-planning platforms and personalized nutrition apps to meet evolving consumer needs. By leveraging data analytics, I aim to empower individuals to make informed dietary decisions and drive positive changes in the food industry.

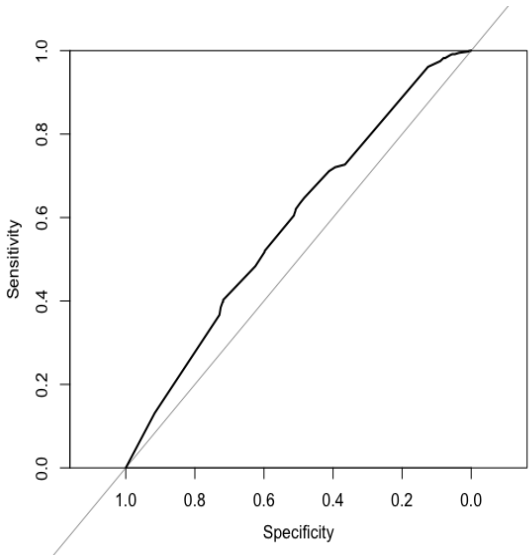
Appendices:

Random Forest:

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 349 275
         1 499 677

               Accuracy : 0.57
                 95% CI : (0.5468, 0.593)
    No Information Rate : 0.5289
    P-Value [Acc > NIR] : 0.0002531

                  Kappa : 0.1245

 Mcnemar's Test P-Value : 1.096e-15

            Sensitivity : 0.7111
            Specificity : 0.4116
         Pos Pred Value : 0.5757
         Neg Pred Value : 0.5593
             Prevalence : 0.5289
         Detection Rate : 0.3761
   Detection Prevalence : 0.6533
      Balanced Accuracy : 0.5613

       'Positive' Class : 1
```





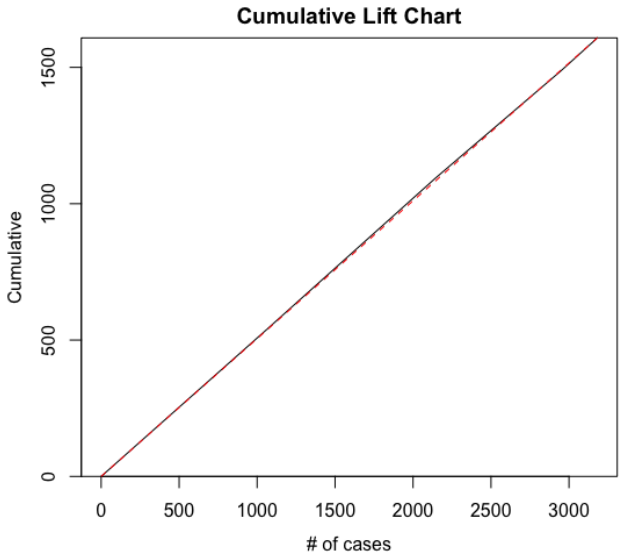


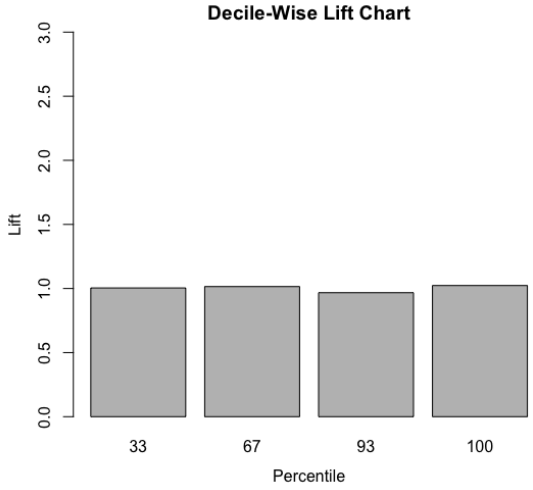
K-Nearest Neighbor:

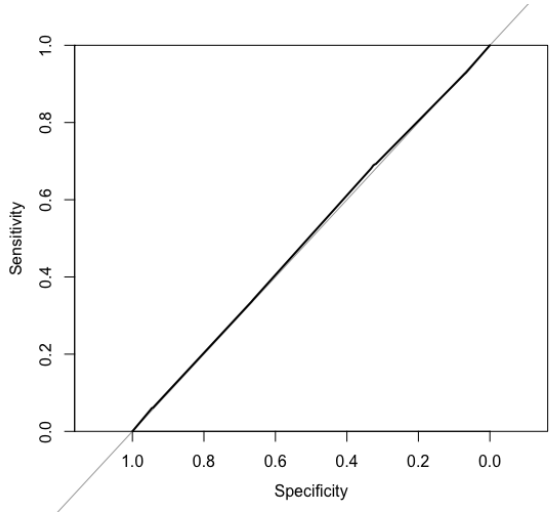
```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 786 775
         1 788 833

               Accuracy : 0.5088
                 95% CI : (0.4913, 0.5263)
    No Information Rate : 0.5053
    P-Value [Acc > NIR] : 0.3549

                  Kappa : 0.0174

 Mcnemar's Test P-Value : 0.7615

            Sensitivity : 0.5180
            Specificity : 0.4994
         Pos Pred Value : 0.5139
         Neg Pred Value : 0.5035
             Prevalence : 0.5053
         Detection Rate : 0.2618
   Detection Prevalence : 0.5094
      Balanced Accuracy : 0.5087

       'Positive' Class : 1
```

```
k-Nearest Neighbors

4776 samples
  10 predictor
   2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 4299, 4299, 4299, 4298, 4298, 4298, ...
Resampling results across tuning parameters:

  k   Accuracy   Kappa
   1  0.5144466  0.02889594
   2  0.5083770  0.01683177
   3  0.5161193  0.03226838
   4  0.5226113  0.04519470
   5  0.5226104  0.04510000
   6  0.5106747  0.02099116
   7  0.5127782  0.02527457
   8  0.5121545  0.02399096
   9  0.5148772  0.02942186
  10  0.5159255  0.03139837

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 4.
```

Depth of File	N	Cume N	Mean Resp	Cume Mean Resp	Cume Pct of Total Resp	Lift Index	Cume Lift	Mean Model Score
33	1064	1064	0.51	0.51	33.6%	100	100	0.79
67	1080	2144	0.51	0.51	68.0%	102	101	0.50
93	804	2948	0.49	0.50	92.5%	97	100	0.26
100	234	3182	0.52	0.51	100.0%	102	100	0.01
NA	NA	NA	NA	NA	NA%	NA	NA	NA
NA	NA	NA	NA	NA	NA%	NA	NA	NA
NA	NA	NA	NA	NA	NA%	NA	NA	NA
NA	NA	NA	NA	NA	NA%	NA	NA	NA
NA	NA	NA	NA	NA	NA%	NA	NA	NA
NA	NA	NA	NA	NA	NA%	NA	NA	NA





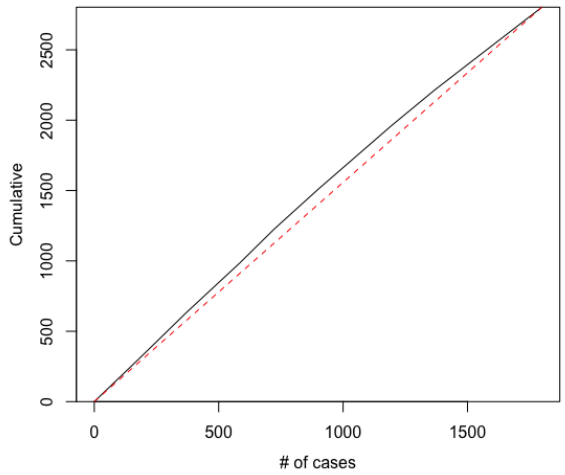


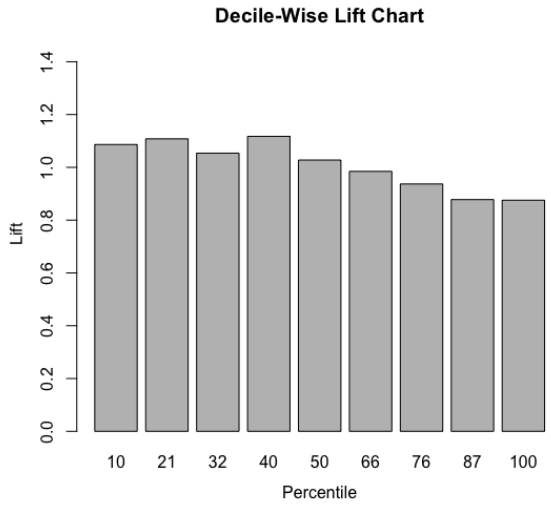
Naive Bayes:

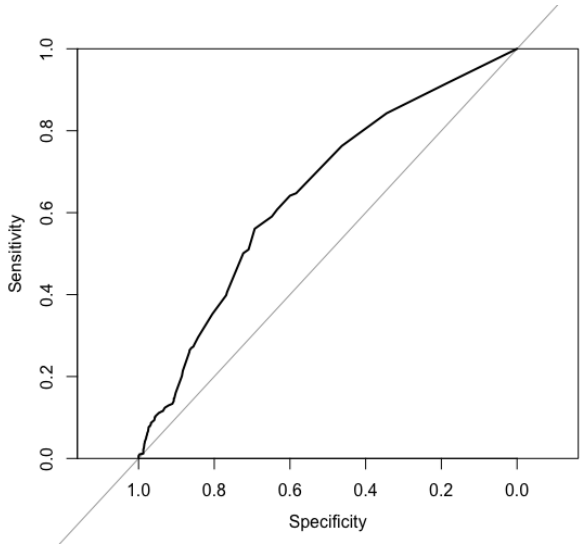
```
Naive Bayes

2701 samples
   7 predictor
   2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 2432, 2431, 2431, 2431, 2430, 2431, ...
Resampling results across tuning parameters:

  usekernel  Accuracy   Kappa
  FALSE      0.6327546  0.2785668
   TRUE      0.5479364  0.1608751

Tuning parameter 'fL' was held constant at a value of 0
Tuning parameter 'adjust' was held constant at
 a value of 1
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were fL = 0, usekernel = FALSE and adjust = 1.
```

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 564 492
         1 231 512

               Accuracy : 0.5981
                 95% CI : (0.575, 0.6209)
    No Information Rate : 0.5581
    P-Value [Acc > NIR] : 0.0003297

                  Kappa : 0.2121

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.5100
            Specificity : 0.7094
         Pos Pred Value : 0.6891
         Neg Pred Value : 0.5341
             Prevalence : 0.5581
         Detection Rate : 0.2846
   Detection Prevalence : 0.4130
      Balanced Accuracy : 0.6097

       'Positive' Class : 1
```

Depth of File	N	Cume N	Mean Resp	Cume Mean Resp	Cume Pct of Total Resp	Lift Index	Cume Lift	Mean Model Score
10	179	179	1.69	1.69	10.8%	109	109	1.00
21	190	369	1.73	1.71	22.5%	111	110	0.93
32	215	584	1.64	1.68	35.1%	105	108	0.69
40	139	723	1.74	1.70	43.7%	112	109	0.55
50	178	901	1.60	1.68	53.9%	103	108	0.48
66	292	1193	1.53	1.64	69.9%	98	105	0.26
76	174	1367	1.46	1.62	79.0%	94	104	0.17
87	204	1571	1.37	1.59	88.9%	88	102	0.15
100	228	1799	1.36	1.56	100.0%	88	100	0.12
NA	NA	NA	NA	NA	NA%	NA	NA	NA



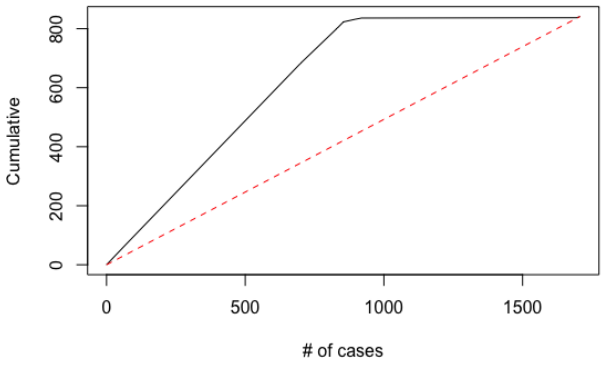


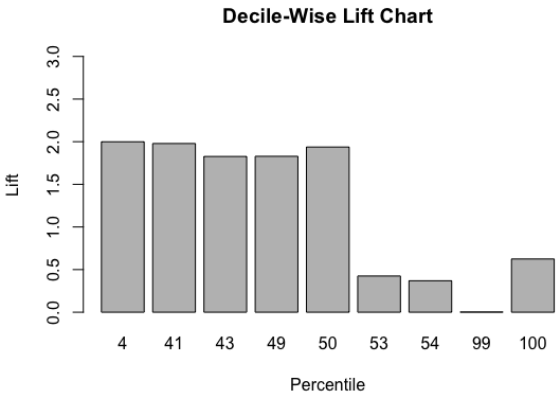


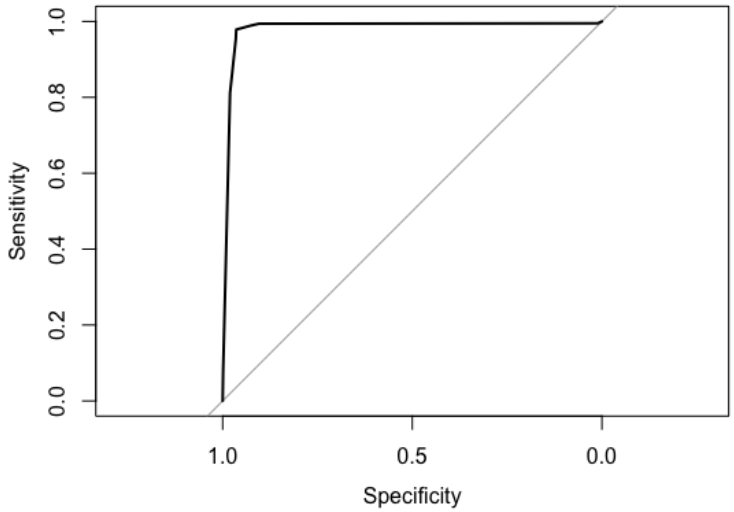
Classification Tree:

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 835  18
         1  31 823

               Accuracy : 0.9713
                 95% CI : (0.9622, 0.9787)
    No Information Rate : 0.5073
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.9426

 Mcnemar's Test P-Value : 0.08648

            Sensitivity : 0.9786
            Specificity : 0.9642
         Pos Pred Value : 0.9637
         Neg Pred Value : 0.9789
             Prevalence : 0.4927
         Detection Rate : 0.4821
   Detection Prevalence : 0.5003
      Balanced Accuracy : 0.9714

       'Positive' Class : 1
```

Depth of File	N	Cume N	Mean Resp	Cume Mean Resp	Cume Pct of Total Resp	Lift Index	Cume Lift	Mean Model Score
4	66	66	0.98	0.98	7.7%	200	200	0.98
41	635	701	0.97	0.98	81.3%	198	198	0.98
43	30	731	0.90	0.97	84.5%	183	197	0.96
49	101	832	0.90	0.96	95.4%	183	196	0.92
50	22	854	0.95	0.96	97.9%	194	196	0.84
53	43	897	0.21	0.93	98.9%	42	188	0.11
54	22	919	0.18	0.91	99.4%	37	185	0.09
99	775	1694	0.00	0.49	99.5%	0	100	0.00
100	13	1707	0.31	0.49	100.0%	62	100	0.00


Cumulative
0
200
400
600
800
0
500
1000
1500
# of cases


Decile-Wise Lift Chart
Lift
0.0
0.5
1.0
1.5
2.0
2.5
3.0
4
41
43
49
50
53
54
99
100
Percentile


Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
1.0
0.5
0.0
Specificity