<FINDING OPTIMAL CLASSIFIER>

KOLBE WILLIAMS

ADEN TESSMAN

THE DATASET:
CARD TRANSACTIONS

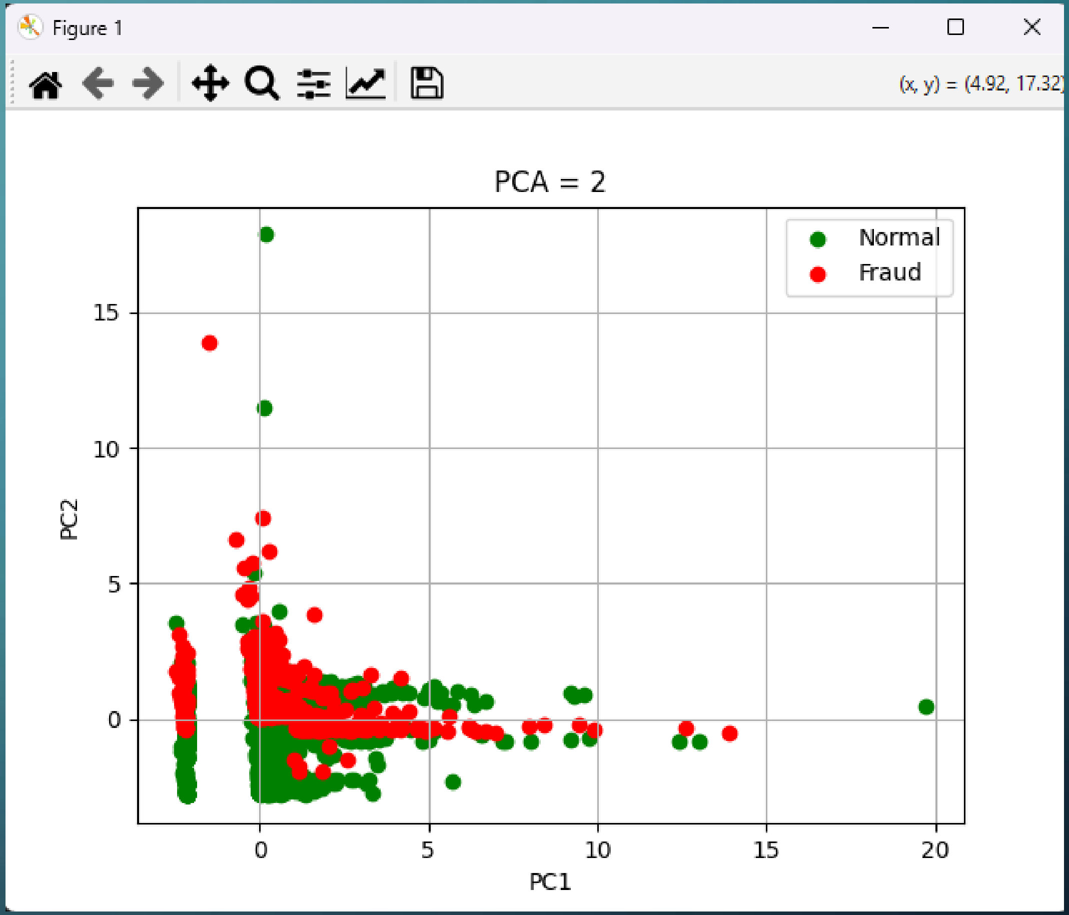
Total Features	• 7
Continuous Features	• distance_from_home, distance_from_last_transaction, ratio_to_median_purchase_price
Binary Features	• repeat_retailer, used_chip, used_pin_number, online_order
Classes	• Fraudulent (0 or 1)

CLASSIFICATION MODELS USED

- KNN
  - non – linear classifier (and hence, the prediction boundary is non-linear) that predicts which class a new test data point belongs to by identifying its k nearest neighbors' class. Which may benefit from our dataset
- Logistic Regression
  - linear classification mode which is used to model binary dependent variables. It is used to predict the probability (p) that an event occurs. This may now preform as well due to the mix of binary and continuous values
- Naïve Bayes
  - Popular method based off the bayes theorem, however, assumptions made are that all the features are **independent** of one another and contribute equally to the outcome; all are of **equal** importance. But these assumptions are not always valid
- Random Forest
  - Ensemble learning, where multiple Decision trees are put together to create one bigger and better performing ML algorithm. Decision tree is a flowchart, where each internal node denotes a test on an attribute, each subsequent branch represents an outcome of the test (True or False), and each leaf node (terminal node) holds a class label. Based on this tree, splits are made to differentiate classes in the original dataset given. Powerful and accurate, good performance on many problems, including non – linear.
- SVM
  - Separate the classes by drawing a hyperplane such that there is a maximum margin determined by the hyper parameter c (cost). The lines of the margins are referend to as support vectors. This ML model can be used as a linear or non-linear classifier based on the kernel used allowing for some flexibility..

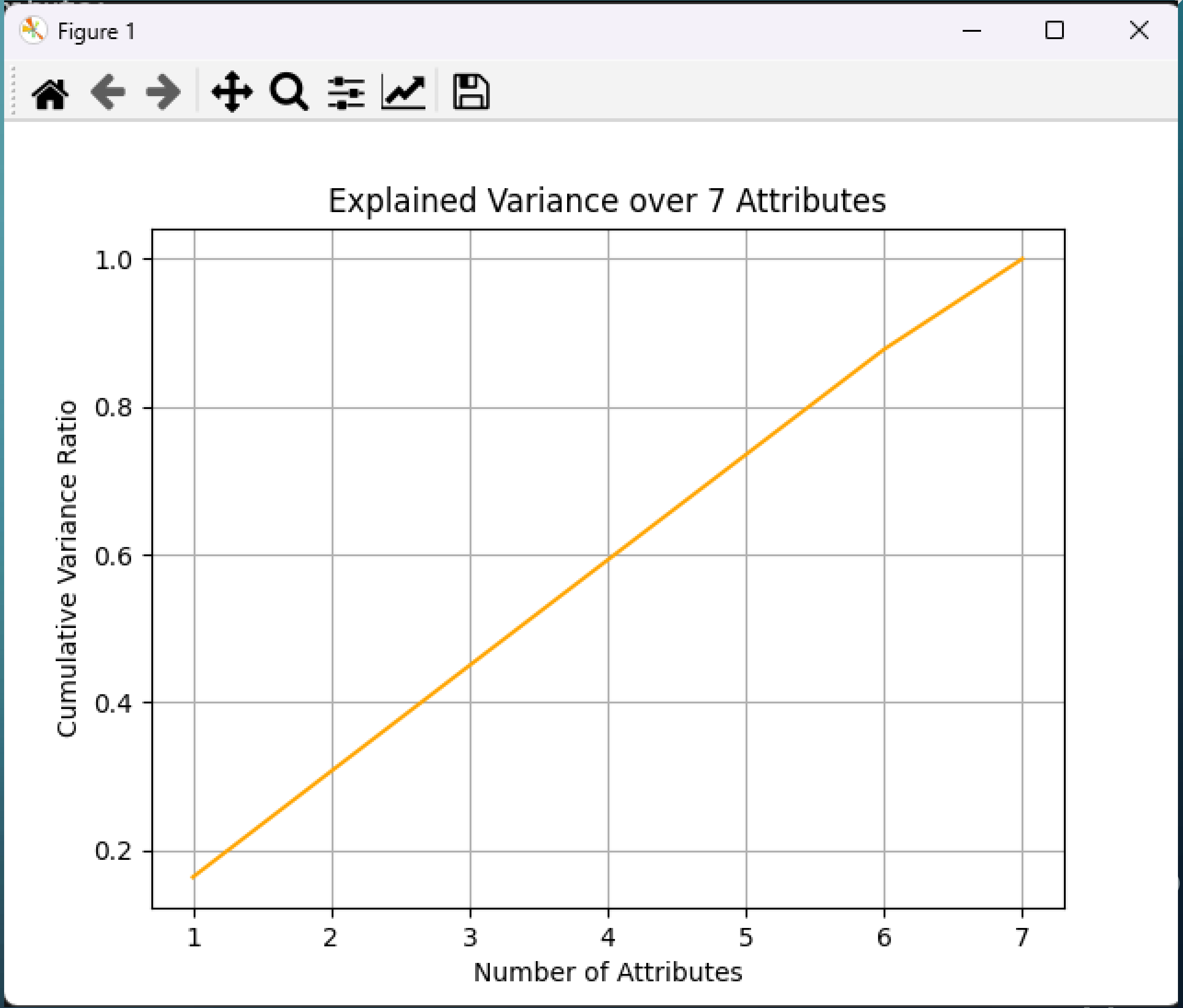
ISSUES WITH PCA

- We attempted utilize PCA to reduce the features so could we visualize trends on a 2D plot, However, PCA was designed to handle continuous value. Due to majority of the features being discrete, it resulted in a plot that did not give insight into the patterns of the classes

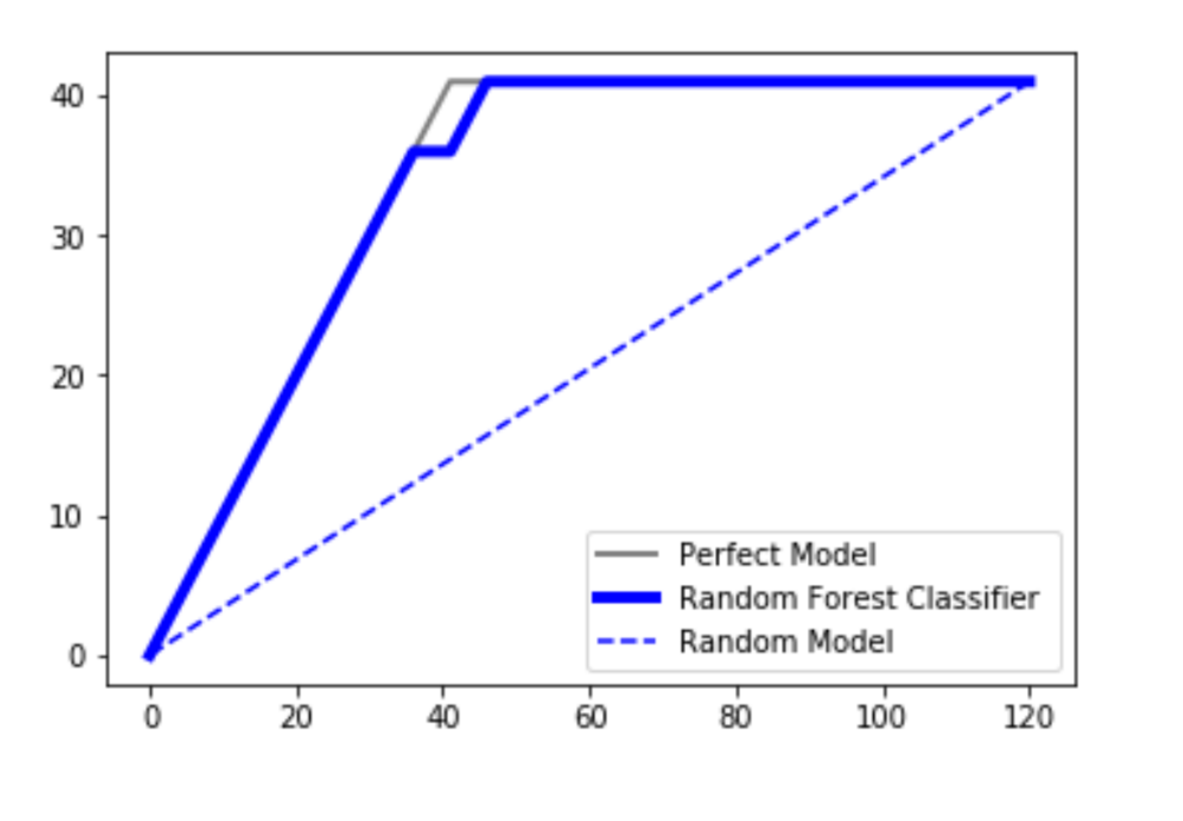


PCA CONTINUED

- However, we did to retrieve the variance ratio of each of the features
- The cumulative variance was linear meaning that our features did not have a lot of variance.
- This is due to 4/7 features being discrete and uncorrelated



VISUALIZATION OF MODEL PERFORMANCE:
CAP PLOTS

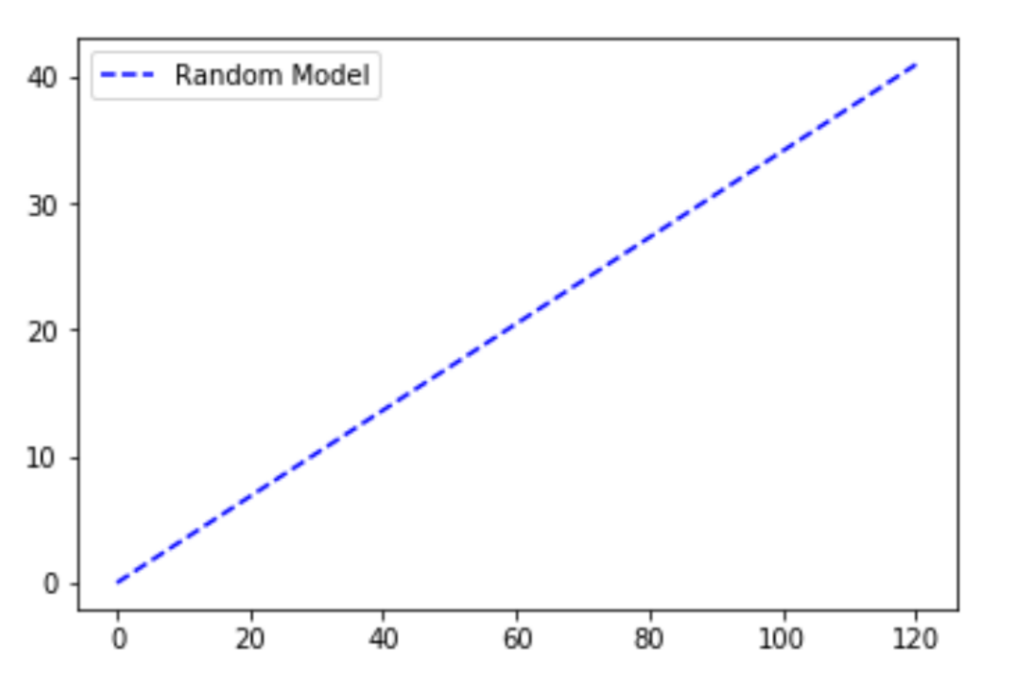


CAP (Cumulative Accuracy Profile) is used in the performance evaluation classification models. It helps to visualize the overall robustness of a model In order to visualize this, three distinct curves are plotted in our plot:

- A random plot
- A plot obtained by the predictions of a ML algorithm against the TP and TN
- A perfect plot(an ideal line)

The CAP of a model represents the cumulative number of positive outcomes along the y-axis versus the corresponding cumulative number of prediction values

VISUALIZATION OF MODEL PERFORMANCE:
CAP PLOTS | X AND Y, RANDOM MODEL

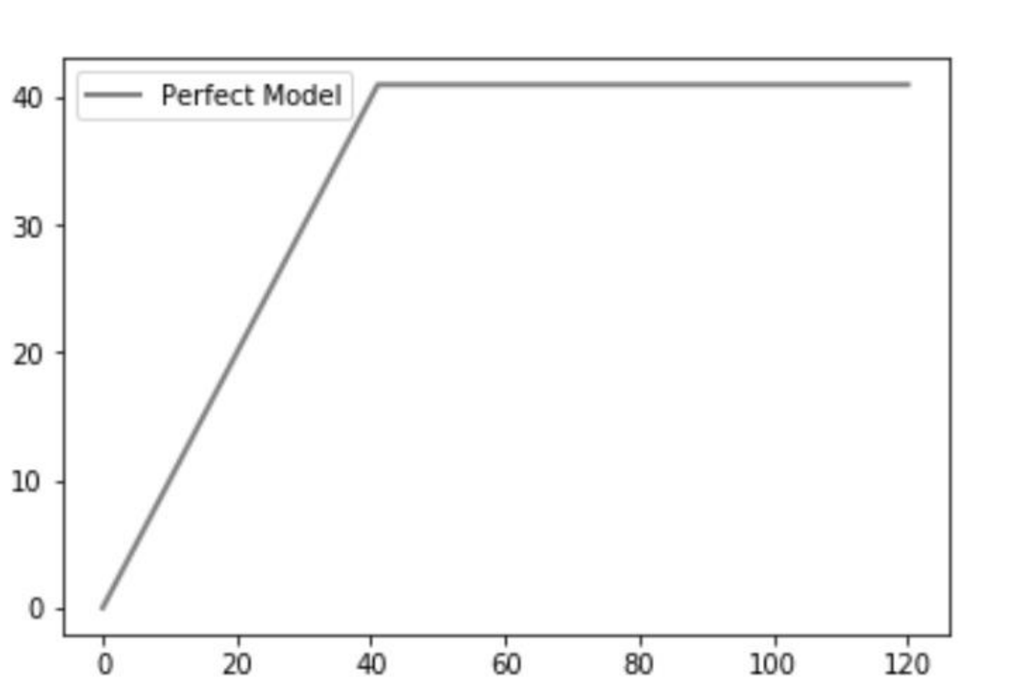


X-axis:

  - total data instances (e.g., card transaction).

- Y-axis:
  - The total positive instances captured
- Random Model Line
  - Represents a model with no predictive power. The performance of a model that chooses true or false randomly

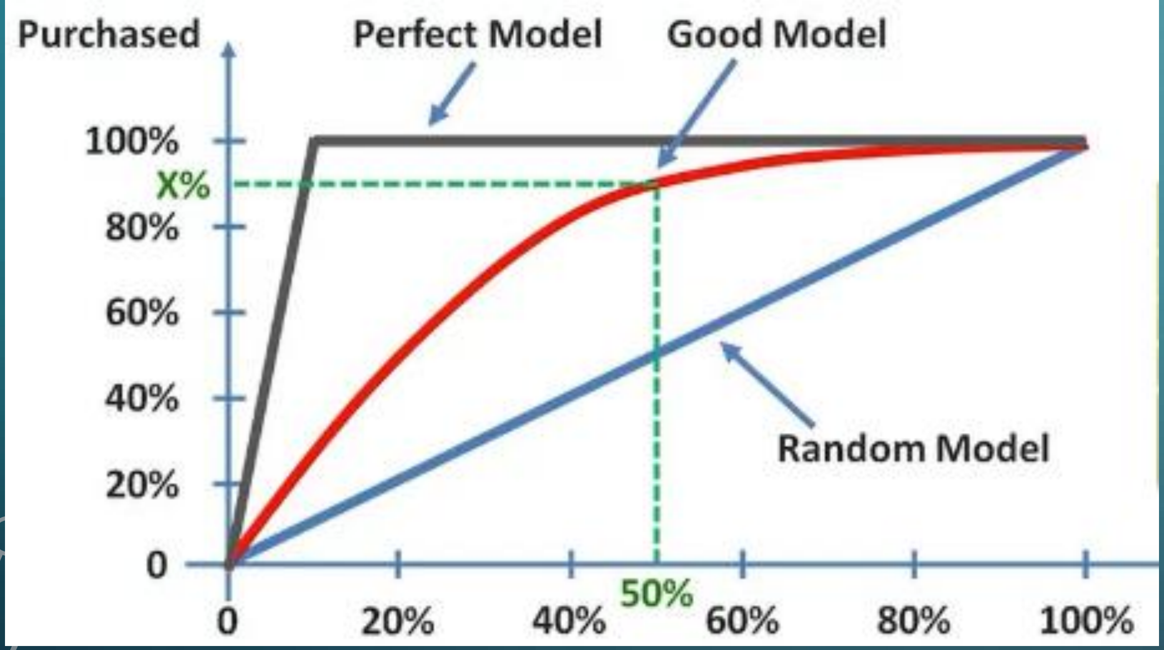
VISUALIZATION OF MODEL PERFORMANCE:
CAP PLOTS | PERFECT MODEL



Perfect Model line:

- Display how your model should plot if no misclassifications are made
- The slope of the line is initially steep due to the nature of CAP graphs
  - CAP initially sorts true labels by value (greatest to least, important to note that the order of this is preserved to correctly group corresponding true labels and predicted labels)
- The closer your model is to this line the better

VISUALIZATION OF MODEL PERFORMANCE:
MODEL LINE INTERPRETATION



What the Curve Shape Tells You:

Steep Upward Movement

- **Means:** The model is **correctly identifying many TPs** early — most of classifications the model thinks are positives **actually are.**
- **Implication:** Good model performance

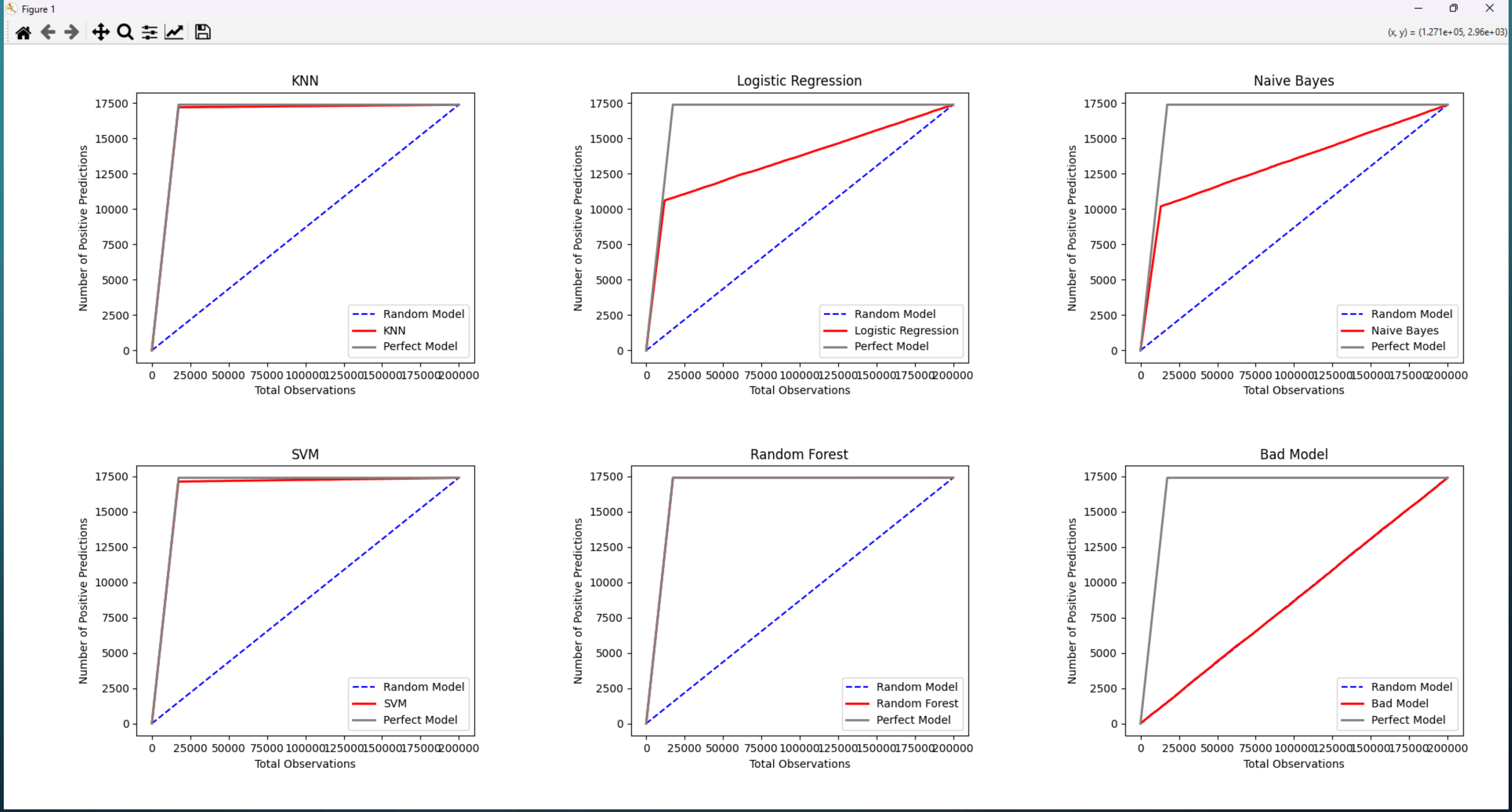
Flat or Shallow Slope

- **Means:** The model is ranking **negatives (false positives)** higher, or missing actual positives (false negatives).
- **Implication:** the model is not distinguishing well in this range.

Plateau at the Top

- Once the curve **reaches 100% on the Y-axis**, that means **all true positives have been found.**
- After that point, you're just going through the rest of the population, which contains **only negatives**, so the curve flattens.

VISUALIZATION OF MODEL PERFORMANCE: OUR GRAPHS



EVALUATION CRITERIA

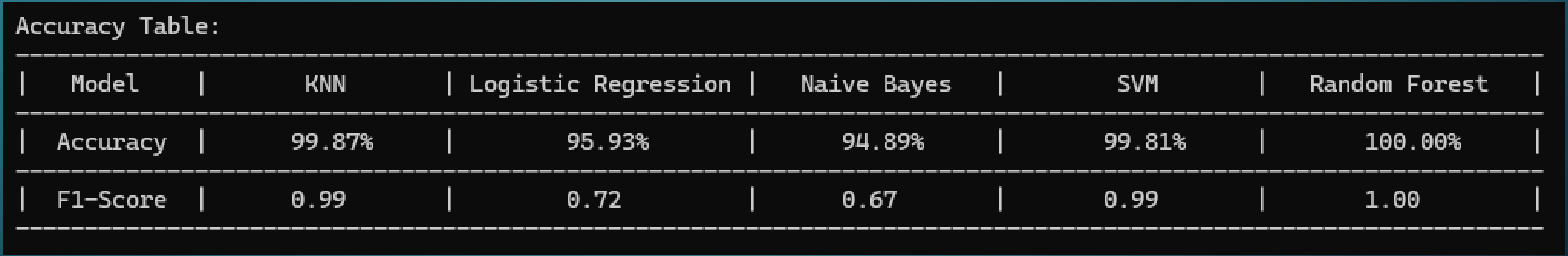
F1 Score

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Evaluation Metric

- We opted to utilize the F1 score to measure overall model performance for each model
- F1 score is the harmonic mean of precision and recall. It combines these metrics into a single value, which allows for a more balanced view of a model performance

RESULTS



```
Accuracy Table:
---------------------------------------------------------------------------------------------------------------
|   Model    |        KNN        | Logistic Regression |    Naive Bayes    |        SVM        |   Random Forest   |
---------------------------------------------------------------------------------------------------------------
|  Accuracy  |       99.87%      |        95.93%       |       94.89%      |       99.81%      |      100.00%      |
---------------------------------------------------------------------------------------------------------------
|  F1-Score  |       0.99        |        0.72         |       0.67        |       0.99        |      1.00         |
---------------------------------------------------------------------------------------------------------------
```

RESULTS CONTINUED

KNN

- Overall KNN preformed very well
- accuracy score of 99.87
- F1 Score of 0.99
  - Shows it was effective at identifying both classes

Logistic Regression

- LogReg preformed relatively poorly
- Decent accuracy score of 95.93
- Poor F1 Score of 0.72
  - Shows it was having issues identifying the minority class (fraud)

Naïve Bayes

- Bayes preformed the worst of all models tested
- Decent accuracy score of 94.89
- Worst F1 Score of 0.67
  - Shows it was not correctly identifying the minority class (fraud) often

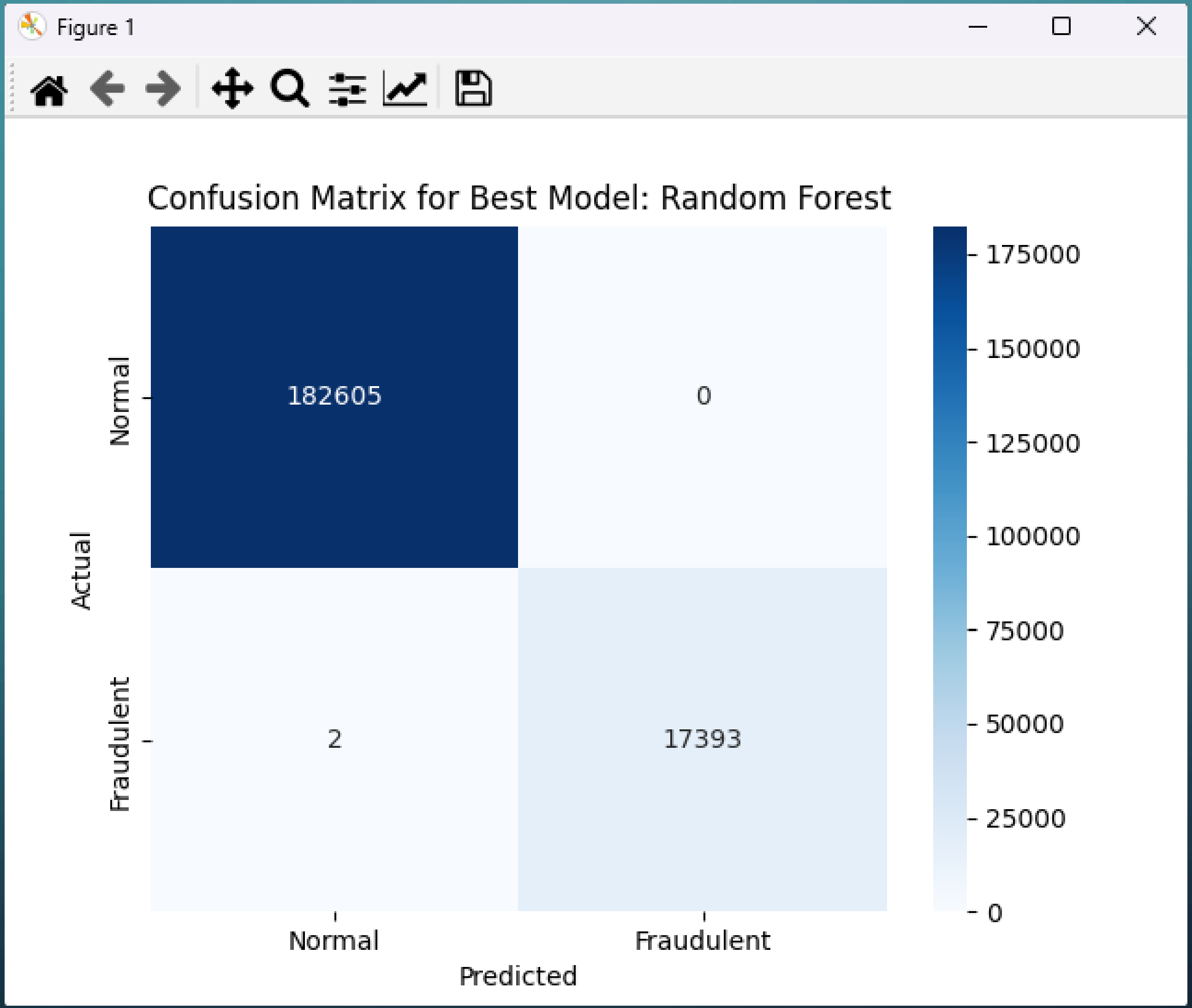
SVM

- Overall SVM preformed very well but sightly worse than KNN
- accuracy score of 99.81
- F1 Score of 0.99
  - Shows it was effective at identifying both classes

Random Forest

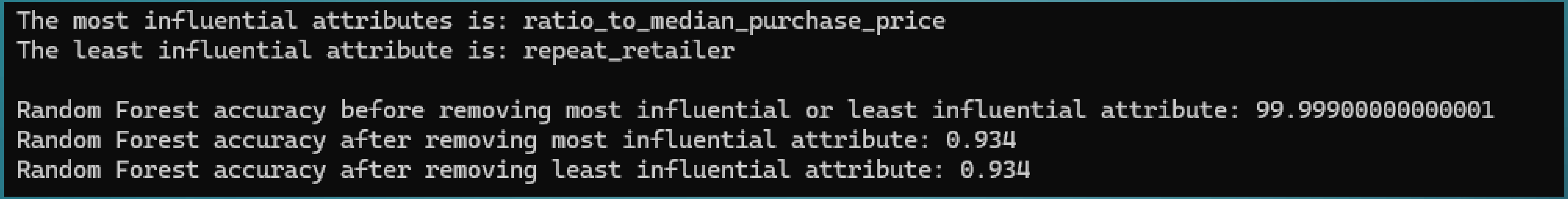
- Best performing model
- accuracy score of 99.999...
- F1 Score of 100
- Very effective with only 2 misclassifications'

CONFUSION MATRIX FOR BEST ALGORITHM

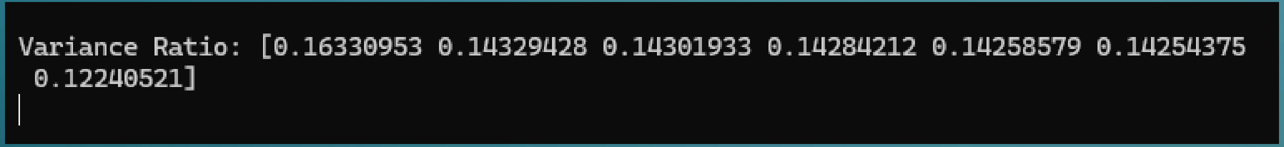


Actual \ Predicted	Normal	Fraudulent
Normal	182605	0
Fraudulent	2	17393

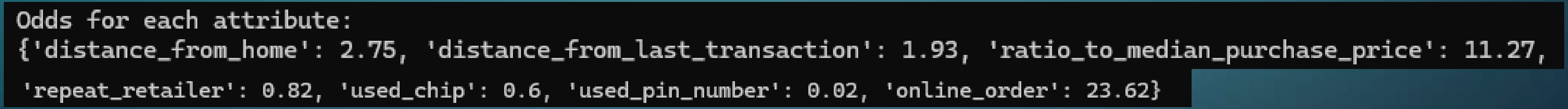
ADDITIONAL FINDINGS



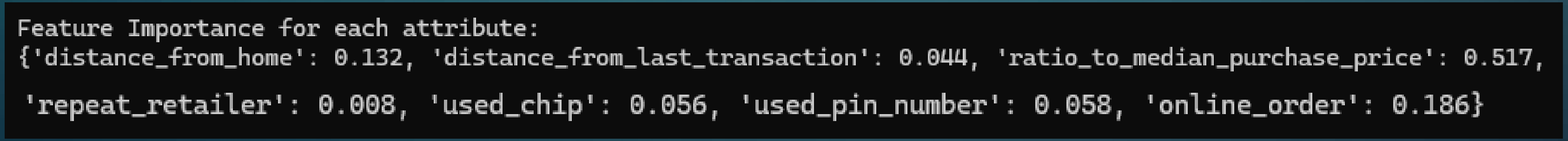
```
The most influential attributes is: ratio_to_median_purchase_price
The least influential attribute is: repeat_retailer

Random Forest accuracy before removing most influential or least influential attribute: 99.99900000000001
Random Forest accuracy after removing most influential attribute: 0.934
Random Forest accuracy after removing least influential attribute: 0.934
```



```
Variance Ratio: [0.16330953 0.14329428 0.14301933 0.14284212 0.14258579 0.14254375
 0.12240521]
```



```
Odds for each attribute:
{'distance_from_home': 2.75, 'distance_from_last_transaction': 1.93, 'ratio_to_median_purchase_price': 11.27,
 'repeat_retailer': 0.82, 'used_chip': 0.6, 'used_pin_number': 0.02, 'online_order': 23.62}
```



```
Feature Importance for each attribute:
{'distance_from_home': 0.132, 'distance_from_last_transaction': 0.044, 'ratio_to_median_purchase_price': 0.517,
 'repeat_retailer': 0.008, 'used_chip': 0.056, 'used_pin_number': 0.058, 'online_order': 0.186}
```