Prediction of The Likely Appearance Of a Shark

Machine Learning Formative Report

By

Otuokere Ugochukwu Charles.

Student Number: 3095760

1.0 Introduction

This report will follow the CRISP-DM framework of solving a data analytics problem by helping the Sydney city council predict the likely appearance of a shark at their beach due to recent attacks.

I will be going through five stages of the CRISP-DM Framework

- **Business Understanding** - Defining the problem we want to solve
- **Data Understanding** - Does the data support the business problem?
- **Data Preparation** - Data cleaning, Prepare for modelling.
- **Modeling** - Find a model that suitably classifies the data.
- **Evaluation** - Check how good the selected model is.

2.0 Business Understanding

Due to the number of recent shark attacks at Australian beaches, the governing body *Sydney City Council* SCC will like to predict whether a shark will appear at their beach or not. This is likely to be a classification machine learning problem.

The data provided tell us if a shark appeared under conditions like average dolphins seen per day, seals seen, diversity of prey e.t.c Machine learning is best suited for this task because any number of these conditions could likely affect the appearance of a shark at the shore or close to the shoreline.

Data mining will enable us to apply different machine learning techniques to generate predictions from the features inputted into the models. These relationships are not clearly seen with the human eye or intuition, hence the need for data mining to better understand the relationships between the likely appearance of a shark and the conditions that may influence the appearance of a shark.

Some of the terminologies that will be referred to in this report are;

Model - A machine learning model is a file that has been trained to recognize certain types of patterns.

Variable - A variable is each column on the dataset. A variable could either be categorical or numerical.

Cross-validation - Splits the data into a given number of folds (usually 5 or 10).

Target Variable - This is the variable whose values are to be modeled and predicted by other variables.
Hyperparameters - Hyperparameters are parameters whose values control the learning process and determine the values of model parameters that a learning algorithm ends up learning.

A good solution is a classification model that suitably predicts the likely appearance of a shark at the beach.

3.0 Data Understanding

The dataset provided contains **456 data points (rows)** and **13 variables (columns)**. The table below summarizes the dataset, the data type of each variable, whether it is to be used for feature selection, and reasons for dropping variables that will not be of use for building the model.

Variables	Data Type/Subtype	Feature selection?	Reasons
murkiness.level	Numeric/continuous	**YES**	
avg.dolphins.seen.per.day	Numeric/continuous	**YES**	
diversity.of.prey	Numeric/continuous	**YES**	
seals.seen	Numeric/continuous	**YES**	
water.temp	Numeric/continuous	**NO**	Values all fall within the same range and has a low correlation to target variable.
people.previous.day	Numeric/discrete	**NO**	Scatter plot does not show any relationship to the target variable.
seagull.density	Numeric/continuous	**YES**	
common.surfboard.color	Categorical/nominal	**NO**	This variable has no correlation to the target variable.

overcast.weather	Numeric/Discrete	**NO**	Distribution widget shows it has large majority values as 0. This is not suitable for the model.
time.of.day	Categorical/nominal	**NO**	The rank widget shows a very low correlation of the time of day to the target variable.
previous.week.fishing	Numeric/continuous	**YES**	
shark	Numeric/Discrete	**YES**	Target variable
id	Numeric/Discrete	**NO**	These are unique values

Using the distribution widget in orange, the variables were plotted to better understand the distribution.
Also, the rank widget was used to score variables according to their correlation with the target variable (shark). The rank widget was used for feature selection.

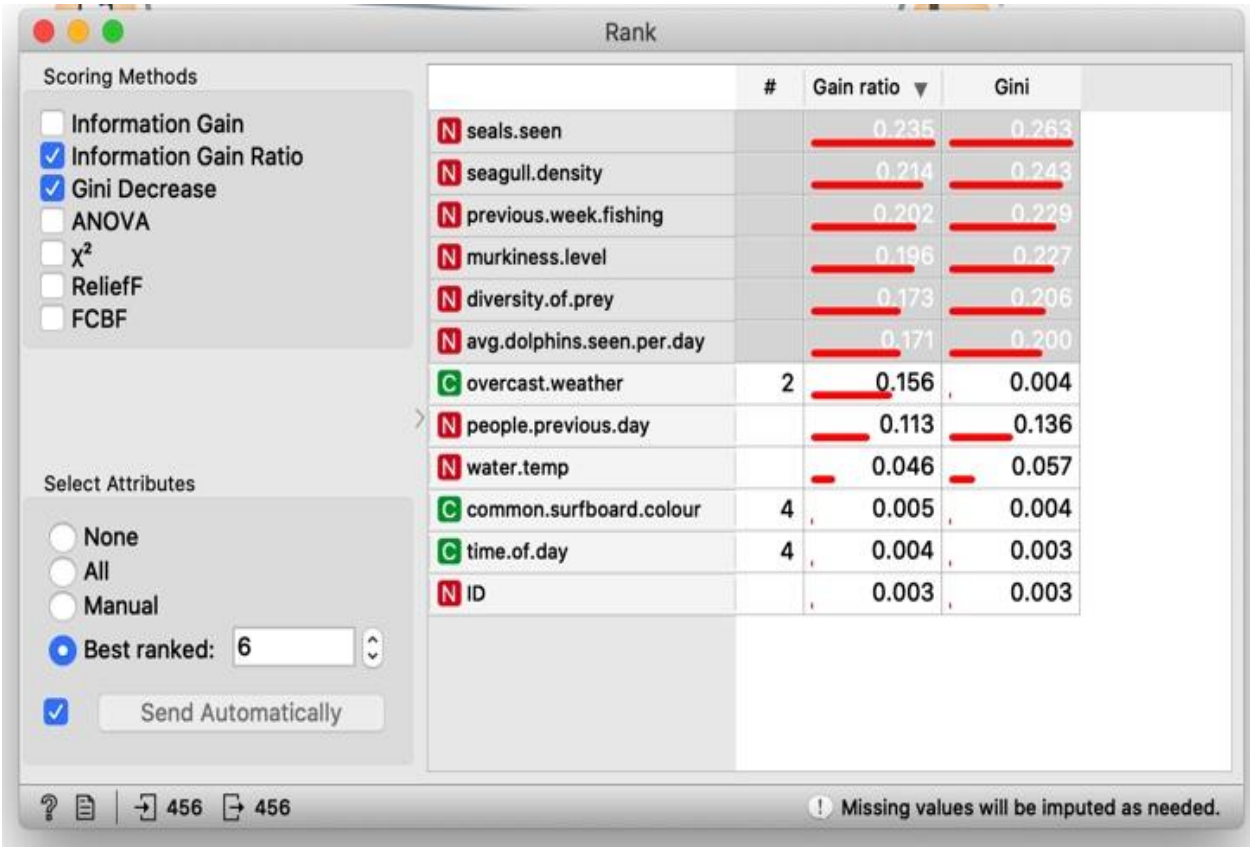


Figure 1: Rank widget showing top 6 variables for feature selection

The best-ranked 6 variables will be used as features for this model as they strongly correlate to the target variable in question.

4.0 Data Preparation

First thing done at this stage was to separate the test and training set. 70% of the data was used for training while 30% will be used later for testing the data.
At this stage, I have selected 6 variables based on the rank score;

- murkiness.level
- avg.dolphins.seen.per.day
- diversity.of.prey
- seals.seen
- seagull.density
- previous.week.fishing

These variables will be used as features to build our model to predict the target variable (shark).

The **feature statistics** widget was used as a summary to figure out datasets with missing values and outliers.

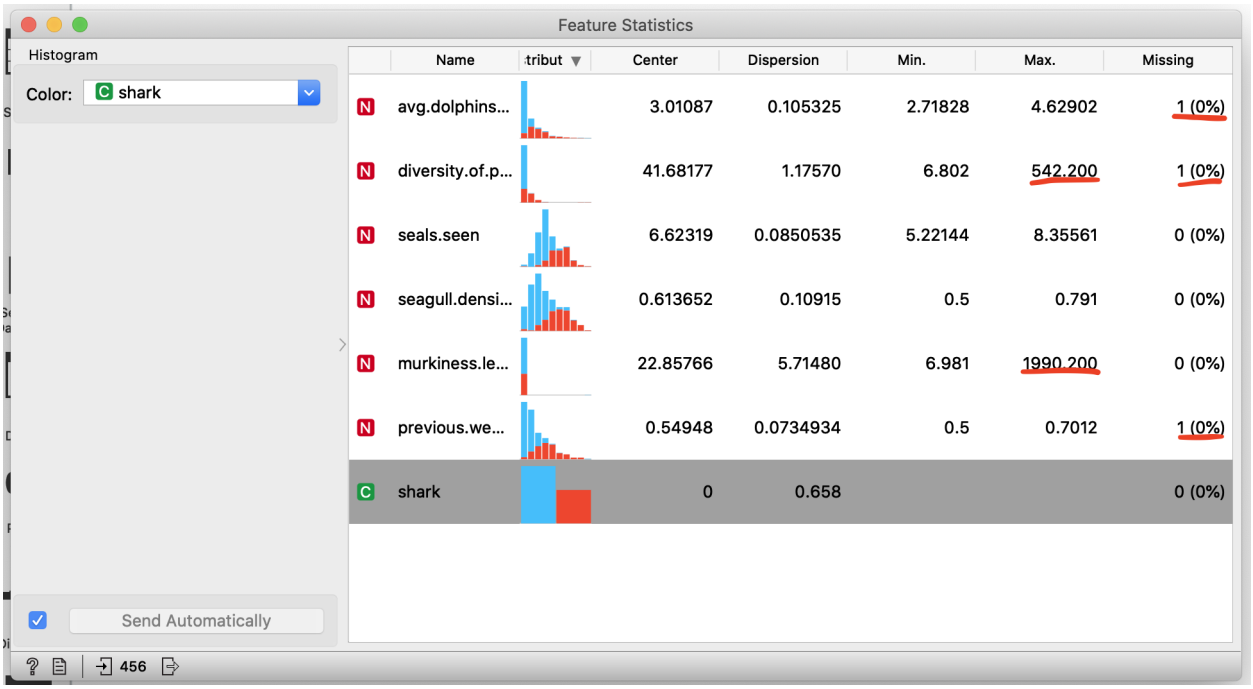


Figure 2: Feature statistics widget of selected variables.

Using the **feature statistics** and **distribution widget** summary, Data Cleaning tasks on selected variables are highlighted below;

Average.dolphins.seen.per.day

- This variable has a missing value which I replaced with the mean of the column by using preprocess widget ***impute missing values preprocessor***.

diversity.of.prey

- This variable has a missing value which I replaced with the mean of the column by using preprocess widget ***impute missing values preprocessor***..
- Also, with a mean of 41, we have a maximum value of 542.20 which I will classify as an outlier.
- Using the Distribution widget to further understand the disparity between the mean value and the max value, we can see that we have two outliers highlighted below.

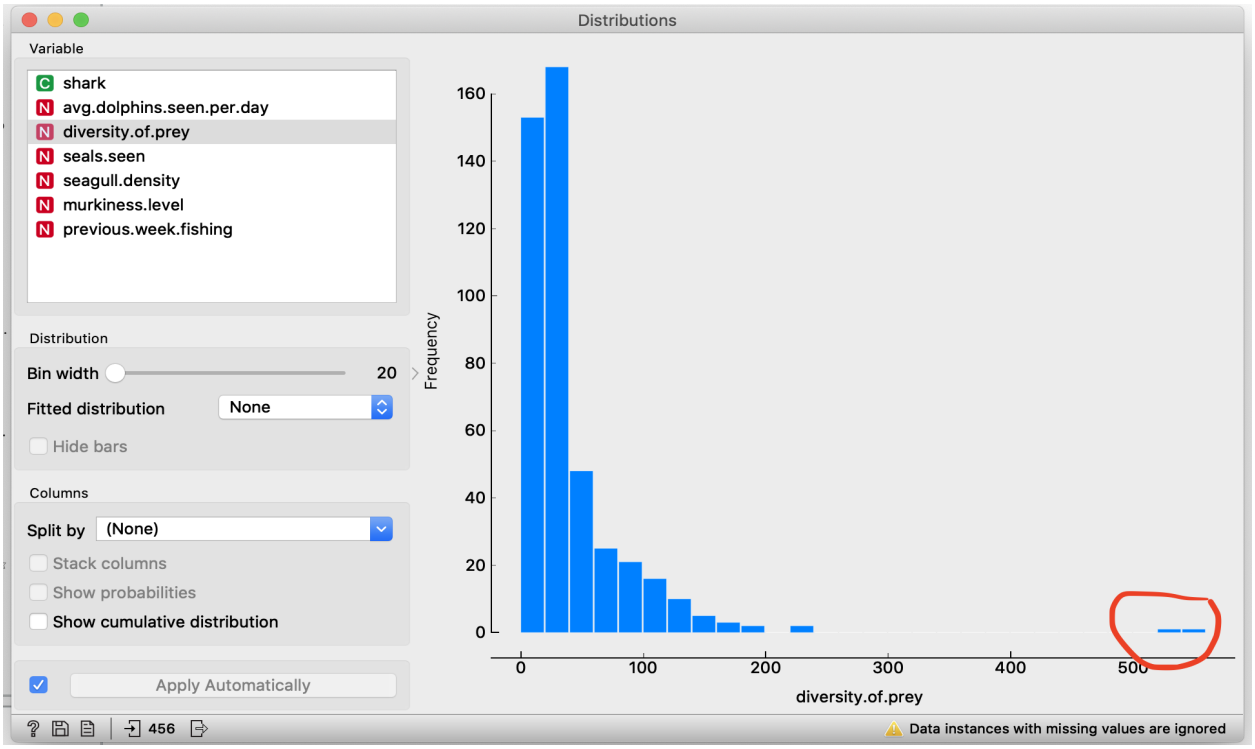


Figure 3: Distributions of the diversity.of.prey variable

- These values were taken out of the model using the select rows widget to filter for **diversity.of.prey** values less than 300.

murkiness.level

- This variable has a mean or center of 22.86 and a maximum value of 1990.200.
- Using the distribution widget, the two outliers were identified below.

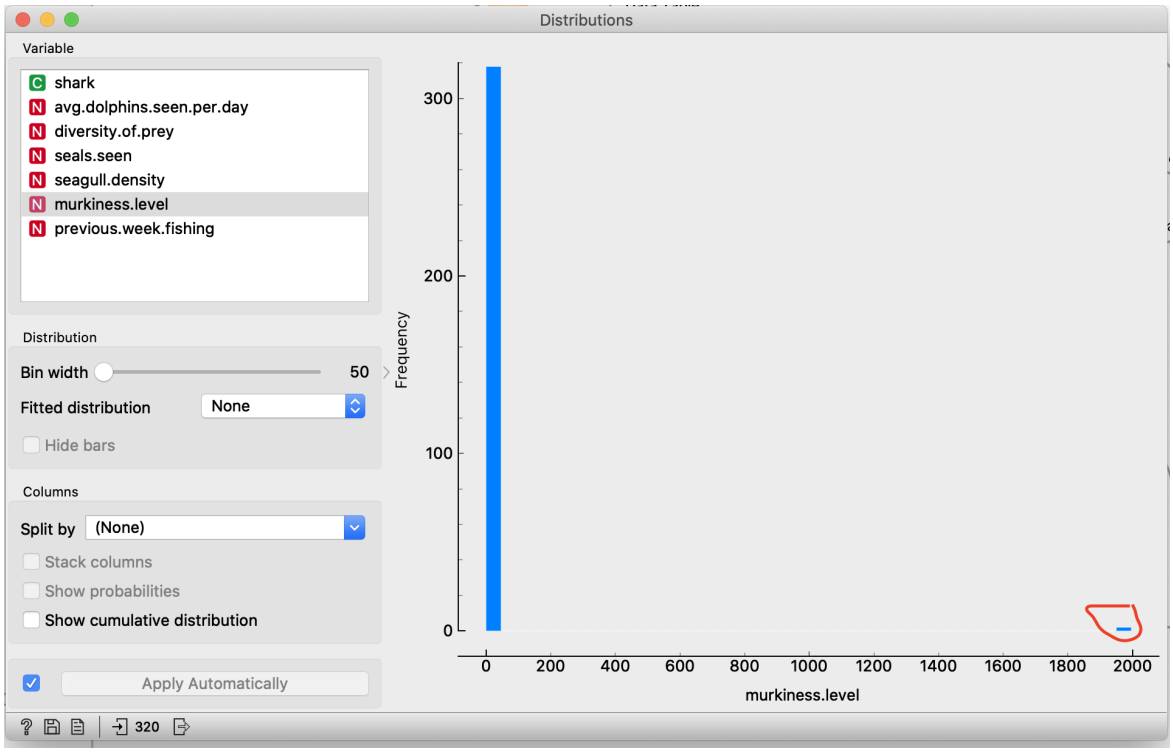


Figure 4 : Distributions of the murkiness.level variable

- These values were taken out of the model using the select rows widget to filter for **murkiness.level** values less than 200.

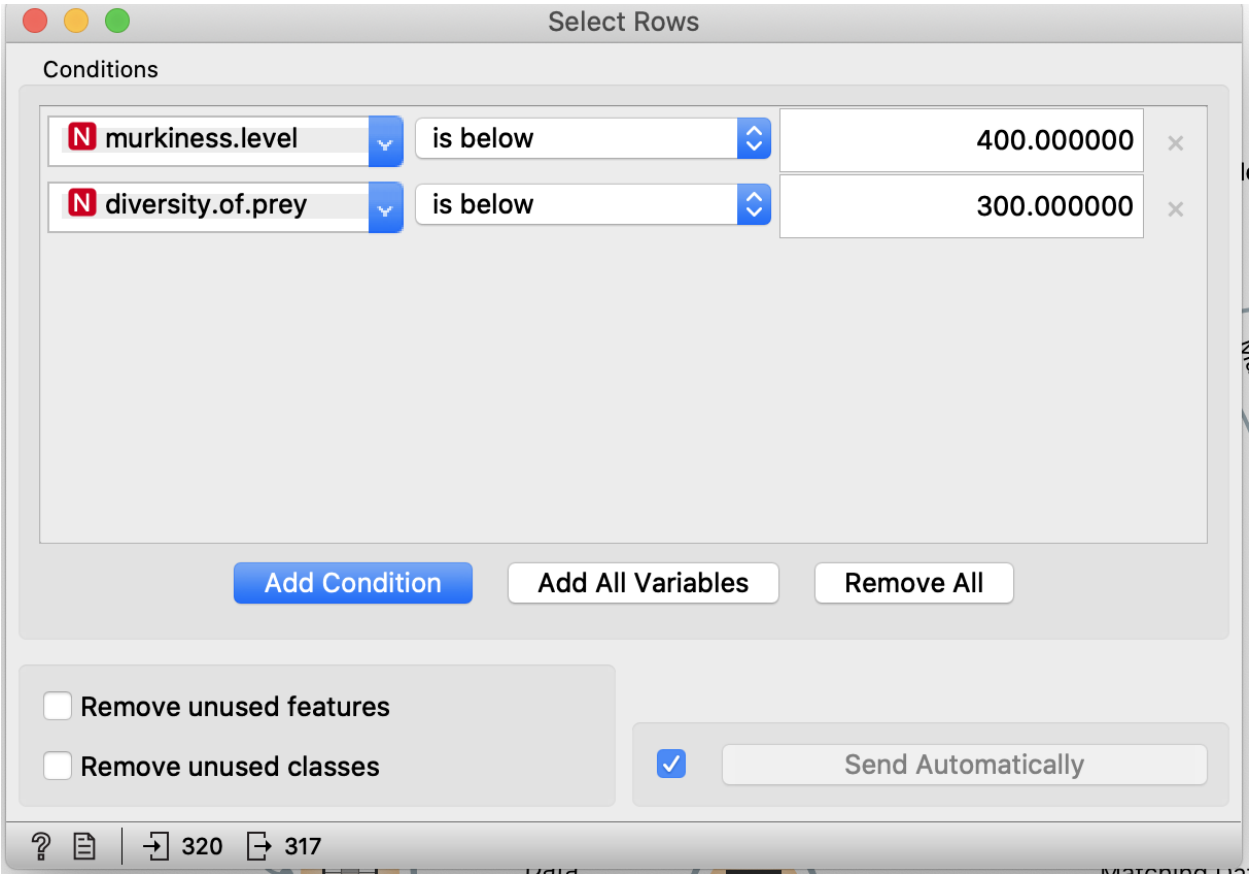


Figure 5: Select row widget to clean the diversity.of.prey and murkiness.level variable column.

previous.week.fishing

- This variable has a missing value which I replaced with the mean of the column by using preprocess widget ***impute missing values preprocessor***.

5.0 Modelling

Modelling of the dataset was done using three different machine learning techniques; **Logistic Regression**, **Random Forest,** and **Neural Networks** to find out which of the models predicts more instances of the appearance of the target variable (shark).

- The data set was split 70% for training and 30% for testing.
- Cross-validation with 5 folds was applied to the training data to train the model.
- Cost function selection (Classification accuracy & Precision)
- Tuning of hyperparameters

The figure below shows the score of each model using the training data to the prediction of the target variable (Shark).

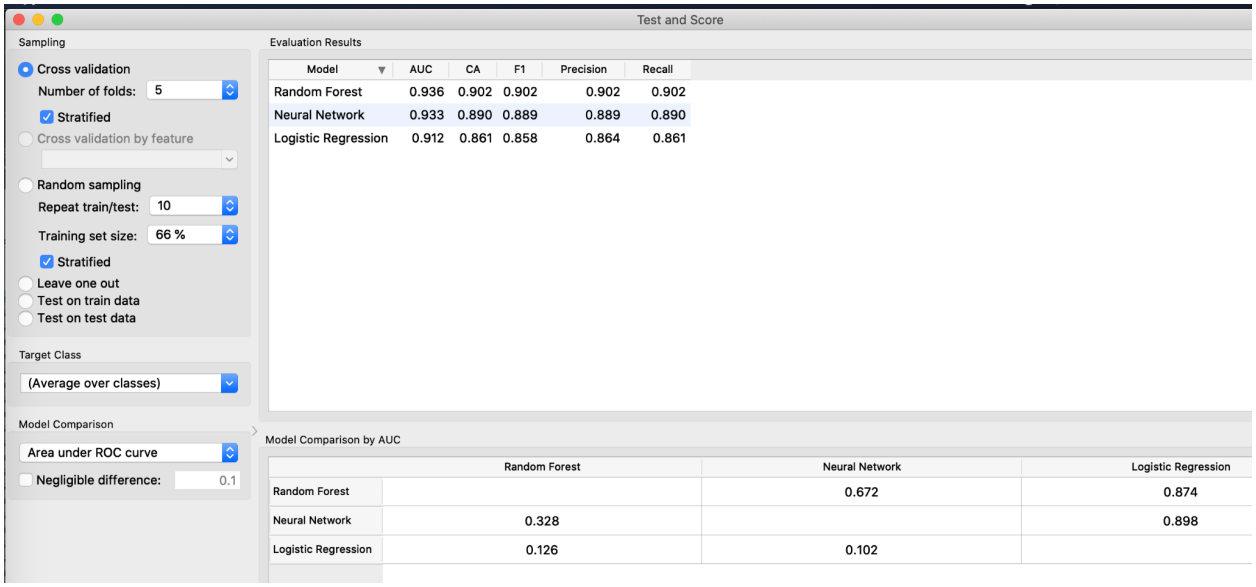


Figure 6: Test and score widget of 3 models with 5-fold cross validation.

For this binary classification task, **Classification accuracy** and **Precision** are used as the cost function to measure how accurate the model performance is.
Classification gives us the proportion of correctly identified examples while Precision gives us the ratio of true positives against predicted positives. Since the appearance of a shark is harmful to humans, it is advantageous to have a high classification accuracy and precision value. One that correctly predicts more instances of a shark appearance based on the training dataset.

Logistic Regression

The logistic regression model achieved a precision of 86%. Tunning the regularization hyperparameter made no significant changes to the precision score. The table below shows different regularization combinations used and the respective score.

Regularization	AUC	CA	F1	PRECISION	RECALL
Lasso(L1)	0.912	0.861	0.858	0.864	0.861
Ridge(L2)	0.935	0.871	0.869	0.871	0.871

The ridge L2 regularization performs better for the linear regression model with classification accuracy and precision of 87%.

Random Forest

The random forest achieved a precision of 90%, which is the highest across the three models. Tuning the hyperparameters number of trees and attributes at each split got the classification accuracy & precision of our model to 90.5%. Not a significant change.

No. of **Trees**	**Attributes at each split**	**AUC**	**CA**	**F1**	**PRECISION**	**RECALL**
25	**5**	**0.929**	**0.905**	**0.905**	**0.905**	**0.905**

Neural Networks

Different hyperparameter tunings were used on this model. We have the details of the best configuration below.

- Neurons in hidden layer = 5
- Activation = Identity
- Solver = SGD
- Regularization = 0.01
- Maximum number of iterations = 200

AUC	CA	F1	PRECISION	RECALL
0.933	0.890	0.889	0.889	0.890

Classification accuracy & Precision are at 89%.

Model Selection

After tunning hyperparameters of each models, we have the following score from the test and score widget in orange.

Model	AUC	CA	F1	PRECISION	RECALL
Logistic Regression	0.935	0.871	0.869	0.871	0.871
Random Forest	0.929	0.905	0.905	0.905	0.905
Neural Networks	0.933	0.890	0.889	0.889	0.890

All 3 trained models performed relatively well in predicting the appearance of a shark.
For this binary classification task, we will use the Random Forest model as the final selected model as it has a higher classification accuracy and precision value.

Also, for testing the different models on the training dataset, the random forest model performs the highest with an accuracy of 96%.

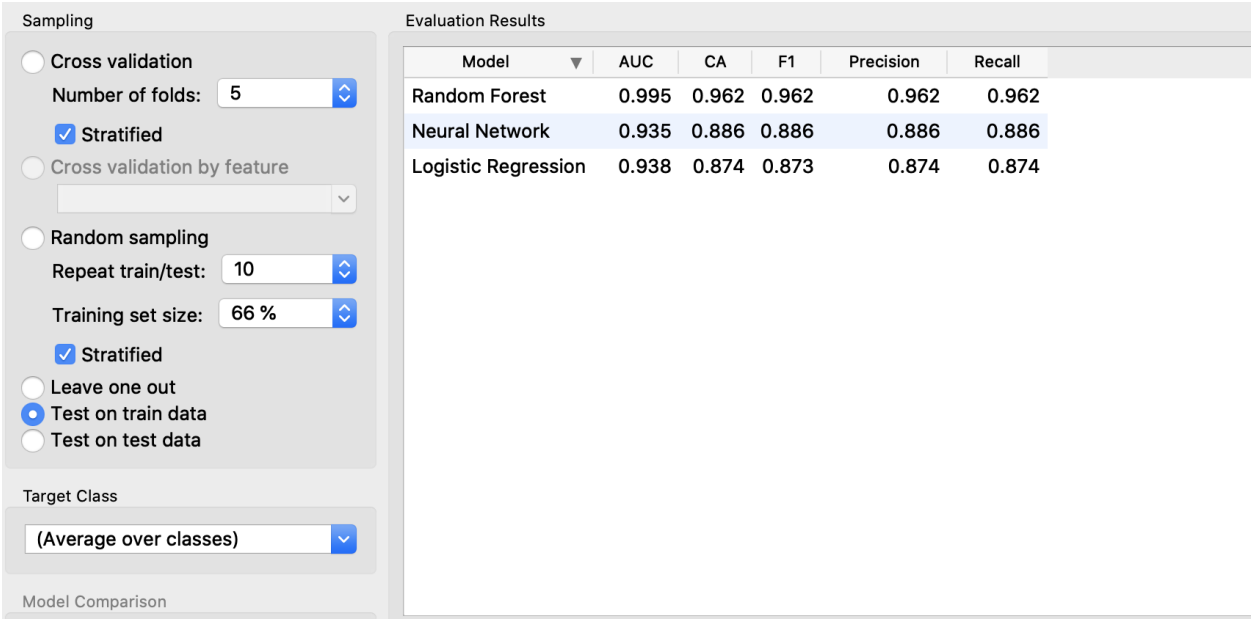
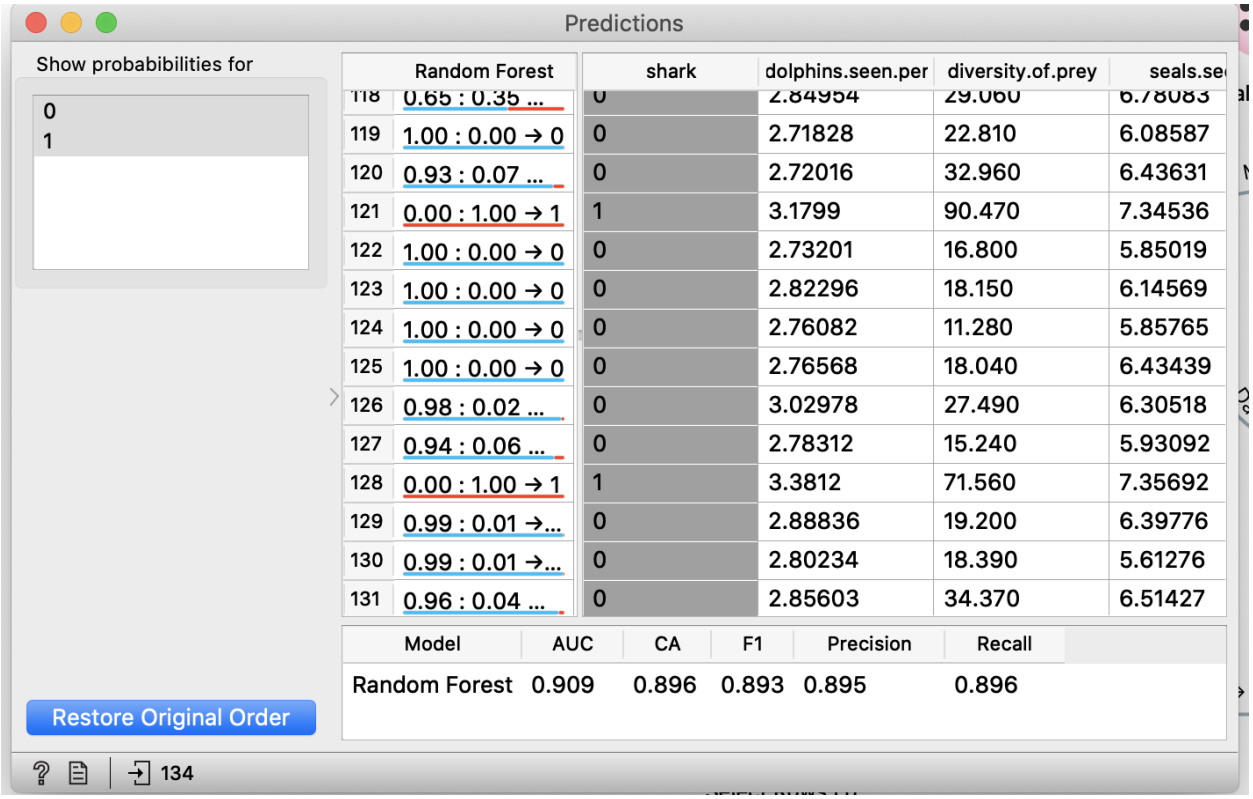


Figure 7: Test and score widget of 3 models with test on train data

6.0 Evaluation

In this final stage of this report, we evaluate the performance of the selected model (random forest) over unseen test data of 134 data points.

The table below shows the summary of model performance gotten from the predictions widget in orange.



	Random Forest	shark	dolphins.seen.per	diversity.of.prey	seals.se
118	0.65 : 0.35 ...	0	2.84954	29.060	6.78083
119	1.00 : 0.00 → 0	0	2.71828	22.810	6.08587
120	0.93 : 0.07 ...	0	2.72016	32.960	6.43631
121	0.00 : 1.00 → 1	1	3.1799	90.470	7.34536
122	1.00 : 0.00 → 0	0	2.73201	16.800	5.85019
123	1.00 : 0.00 → 0	0	2.82296	18.150	6.14569
124	1.00 : 0.00 → 0	0	2.76082	11.280	5.85765
125	1.00 : 0.00 → 0	0	2.76568	18.040	6.43439
126	0.98 : 0.02 ...	0	3.02978	27.490	6.30518
127	0.94 : 0.06 ...	0	2.78312	15.240	5.93092
128	0.00 : 1.00 → 1	1	3.3812	71.560	7.35692
129	0.99 : 0.01 →...	0	2.88836	19.200	6.39776
130	0.99 : 0.01 →...	0	2.80234	18.390	5.61276
131	0.96 : 0.04 ...	0	2.85603	34.370	6.51427

Model	AUC	CA	F1	Precision	Recall
Random Forest	0.909	0.896	0.893	0.895	0.896

Figure 8: Predictions on the selected model

Precision and Classification accuracy of the selected model on the unseen data is at 90% which is an excellent score for our model.

Also, using the confusion matrix to further understand the accuracy of the model in reference to its application;

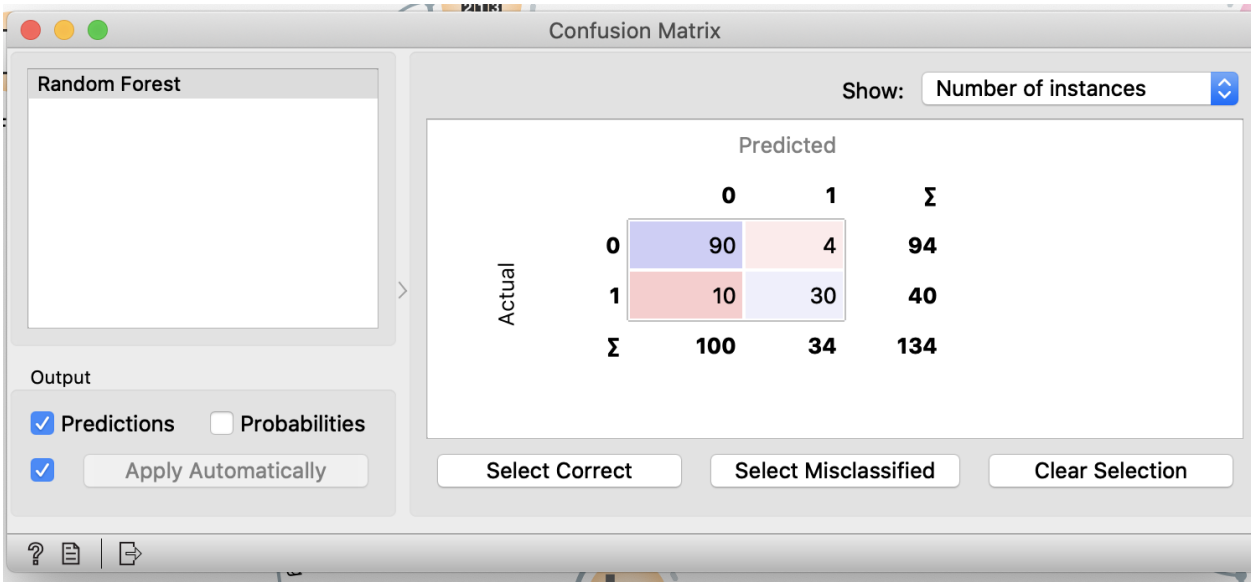
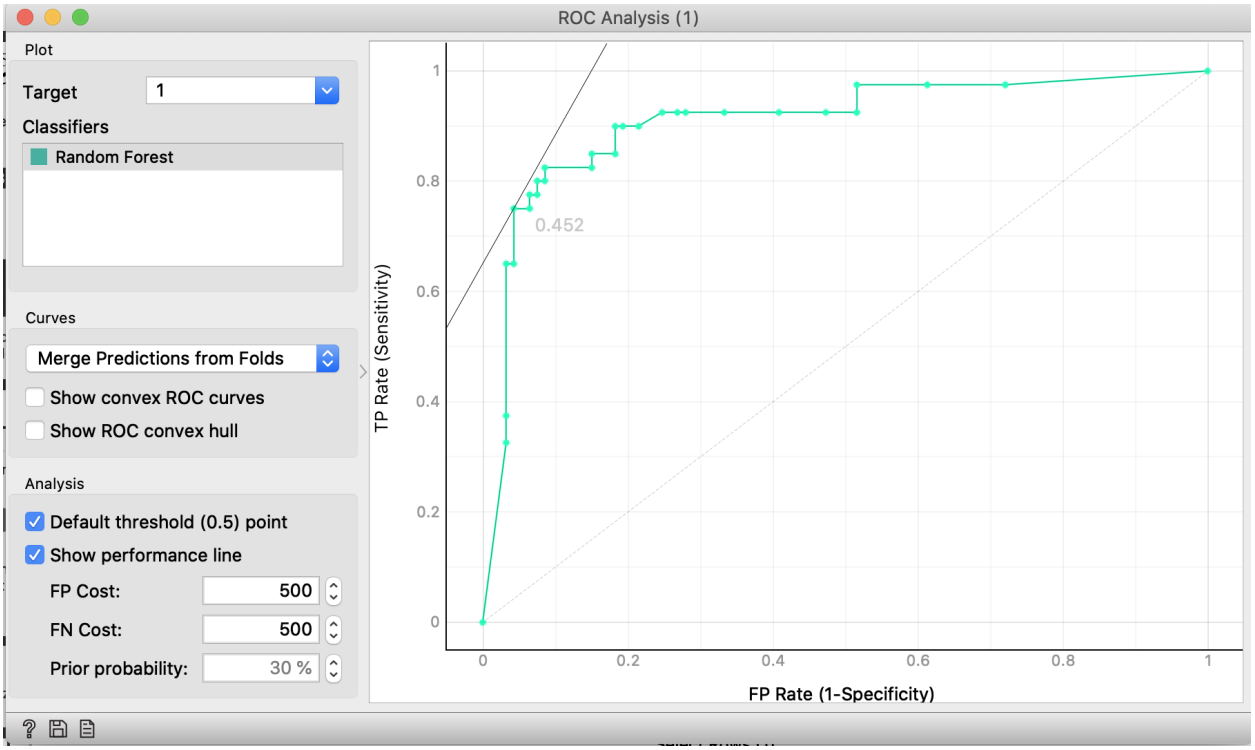


Figure 9: Confusion matrix on model performance.

True Positive (90): For 90 instances where a shark didn't appear, the model correctly predicts no appearance of a shark.
True Negative (30): For 30 instances where a shark did appear, the model correctly predicts the appearance of a shark.
False Negative (10): For 10 instances where a shark appeared, the model wrongly predicts no appearance of a shark. Also called a Type 2 error.
False Positive (4): For 4 instances where a shark didn't appear, the model wrongly predicts the appearance of a shark. Also called a Type 1 error.

For further application of this model, false negatives should be of utmost importance since the Sydney City Council wants to minimize risk to the public, we tune the model to reduce the number of instances of false negatives.
The modelling approach is suitable for the problem of detecting the likely appearance of a shark and the results can be trusted by the company.

Finally, find below the ROC Analysis curve for the selected model. The model performs well for predicting the likely appearance of a shark.


ROC Analysis (1)
Plot
Target
1
Classifiers
Random Forest
Curves
Merge Predictions from Folds
Show convex ROC curves
Show ROC convex hull
Analysis
Default threshold (0.5) point
Show performance line
FP Cost:
500
FN Cost:
500
Prior probability:
30 %
0.452
TP Rate (Sensitivity)
FP Rate (1-Specificity)
0
0.2
0.4
0.6
0.8
1