B39DA – Applied Machine Learning (GA)

Chris Mitchell – H00339901

Table of Contents

Course code and name:	B39DA – Applied Machine Learning (GA)
Type of assessment:	**Individual**
Coursework Title:	Applied Machine Learning Project
Student Name:	Chris Mitchell
Student ID Number:	H00339901

Declaration of authorship. By signing this form:

- **I declare** that the work I have submitted for individual assessment OR the work I have contributed to a group assessment, is entirely my own. I have NOT taken the ideas, writings or inventions of another person and used these as if they were my own. My submission or my contribution to a group submission is expressed in my own words. Any uses made within this work of the ideas, writings or inventions of others, or of any existing sources of information (books, journals, websites, etc.) are properly acknowledged and listed in the references and/or acknowledgements section.

- I confirm that I have read, understood and followed the University's Regulations on plagiarism as published on the University's website, and that I am aware of the penalties that I will face should I not adhere to the University Regulations.

- I confirm that I have read, understood and avoided the different types of plagiarism explained in the University guidance on Academic Integrity and Plagiarism

Student Signature *(type your name):* *Christopher Mitchell*

Date: *27/06/2022*

Copy this page and insert it into your coursework file in front of your title page.
For group assessment, each group member must sign a separate form and all forms must be included with the group submission.

Your work will not be marked if a signed copy of this form is not included with your submission.

Introduction

Organisations of different sizes, successes, and across all sectors face the threat of incidents or risk events. However, for large corporations such as Barclays PLC, the possibility of incidents and risk events is greater and can impact both customers and colleagues immensely. Organisations need to prevent incidents and minimise risks. Therefore, when an incident or risk event does occur, how it is managed is key to ensuring customers and colleagues are not endangered in any way. Organisations must manage their incidents and risks to reduce the number of incidents and risks that occur and give customers and clients a more pleasant experience with the organisation. One of the many departments that make up Barclays PLC is the Chief Controls Office (CCO), the main aim of CCO is to enhance the controls the organisation already has in place and to introduce new controls where there are gaps – helping to reduce incidents and risks to the organisation. However, within CCO a lot of activities focus on governance and reporting type responsibilities, and it is important to both the team that pulls the data together, as well as the stakeholders that receive the data, that there is constant innovation in the governance and reporting space. This is essential as the governance and reporting that is carried out feeds into several meetings and forums throughout the organisation and gets shared at a very senior level. This report will take a sample of incidents and risk events that Barclays have faced, and a scatter plot will be used to give a high-level overview of all the ongoing incidents that the bank has and whether each incident is likely to meet the threshold for a risk event. Following this, a decision tree will be used to inform Barclays whether an incident that the organisation is facing has met the threshold to be classified as a risk event or whether it will remain as an incident.

Problem Statement

Predict whether incidents within Barclays meet the threshold of a risk event, as outlined by Barclays PLC.

To successfully predict whether incidents within Barclays meet the threshold of a risk event, logistic regression and a decision tree will be applied. This will be achieved by firstly extracting data from an internal Barclays system which will then be filtered and labelled to only show incidents and risk events that occurred in the first quarter of

2022. The data will then be closely analysed to ensure the quality of the data is of high quality, with all mandatory fields completed, as the data significantly influences the outcome of the model. Following this, a Scatter Plot will be used as a way of making predictions concerning the results that the model will produce once complete and give an output as to whether each incident within Barclays meets the threshold to be classed as a risk event or not – this will be accomplished by using a supervised learning approach.

Significance

Barclays PLC is a huge international organisation serving around “48 million customers” and clients worldwide (Barclays, No Date). To ensure that the organisation is set up to serve such a vast number of customers, Barclays is made up of thousands of employees spread across the globe, working in different departments and functions to ensure that the operation of the organisation continues to run as effectively and efficiently as possible.

Barclays is one of the largest banks in the UK and was ranked “fifth among the largest banks in Europe” (Statista, 2022). Figure 1 shows the locations Barclays cover in the UK. Many of these locations are branch networks for customers and clients to visit for support and assistance with any of their banking queries or difficulties. Staff in these locations are trained specifically to deal with customers and clients and to support and assist when required.

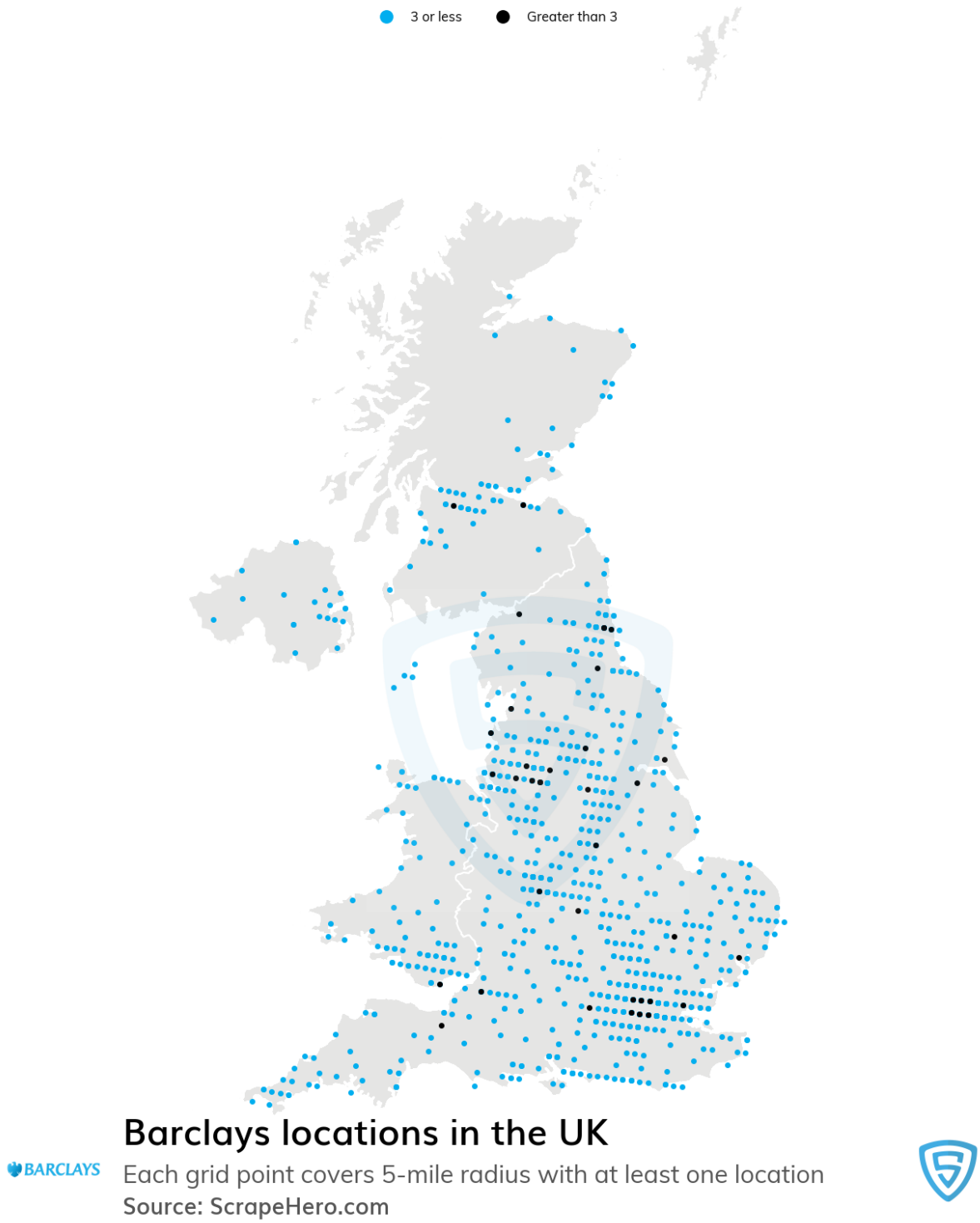


Figure 1: Barclays Locations in the UK (ScrapeHero, 2022)

However, the vast majority of employees within Barclays hold more ‘back office’ based type roles. These roles help to ensure that the organisation is running correctly through roles such as Risk Management, Controls, Technology, and Operations, along with many others. Employees within these types of roles are

based in key Barclays hubs, instead of branch networks. The key Barclays hubs within the UK alone consist of “London, Northampton, Glasgow, Knutsford, Manchester, and Dublin” (Barclays, No Date). These hubs hold many Barclays employees within the UK. However, being separated into hubs can often lead to risk within the organisation because most teams are spread across all locations. These teams can increase the chance of risk since it is harder for teams to communicate in person with those who are separated from the majority of their team and must rely on functions such as phone calls, Microsoft Teams, or emails to have contact with their colleagues. This can lead to low levels of communication within teams and can result in multiple employees duplicating activities that other colleagues have already completed which can result in risk, particularly under circumstances where clients are involved.

Methodology

The Dataset

Although it is the job of CCO to help reduce the number of incidents and risks that the organisation faces - incidents and risks are factors that any operating organisation faces, especially organisations as large as Barclays PLC. This project was designed to make it easier for colleagues and senior management at Barclays to interpret any ongoing incidents and risks that the organisation is currently facing through the use of logistic regression and a decision tree. Logistic regression gives the user an overall view of all the incidents and risk events within the organisation and the decision tree can predict which incidents meet the threshold of a risk event. To do this, the following dataset was used to determine if each incident met the threshold to be classified as a risk event:

Incident_ID	Incident_Title	Reoccuring_Event	Third_Party_Fault	Causal_Area	Impacted_Area	Risk_Event	Rating	Initial_Loss_	Recoveries	Net_Loss	Status	Escalated_Appropriately	Point_Of_Contact
INCIDENT-001	INCIDENT-001 - Title	No	No	Wealth Management	Wealth Management	No	Low	30000	6000	24000	Resolved	Yes	Wealth Management - Incident Handler
INCIDENT-002	INCIDENT-002 - Title	No	Yes	Third Party	Third Party	No	Low	378.99	0	378.99	In Progress	Yes	Third Party - Incident Handler
INCIDENT-003	INCIDENT-003 - Title	No	No	Private Bank	Private Bank	No	Low	21345.12	19546.34	1798.78	In Progress	Yes	Private Bank - Incident Handler
INCIDENT-004	INCIDENT-004 - Title	No	No	Private Bank	Private Bank	No	Low	11008.01	3946.23	7061.78	In Progress	Yes	Private Bank - Incident Handler
INCIDENT-005	INCIDENT-005 - Title	No	Yes	Third Party	Third Party	Yes	High	699000	37000	662000	In Progress	Yes	Third Party - Incident Handler
INCIDENT-006	INCIDENT-006 - Title	No	Yes	Third Party	Third Party	Yes	Low	89824.67	32058.29	57766.38	Under Review	Yes	Third Party - Incident Handler
INCIDENT-007	INCIDENT-007 - Title	No	No	Corporate	Corporate	No	Low	0.1	0.1	0	Under Review	Yes	Corporate - Incident Handler
INCIDENT-008	INCIDENT-008 - Title	Yes	No	Business Management	Business Management	Yes	Medium	333333.33	0	333333.33	Under Review	No	Business Management - Incident Handler
INCIDENT-009	INCIDENT-009 - Title	No	No	Business Management	Business Management	Yes	Medium	499999.99	397679.11	102320.88	Resolved	Yes	Business Management - Incident Handler
INCIDENT-010	INCIDENT-010 - Title	No	No	Premier	Premier	Yes	Medium	101010.93	657.89	100353.04	Resolved	Yes	Premier - Incident Handler
INCIDENT-011	INCIDENT-011 - Title	No	No	Wealth Management	Wealth Management	Yes	Low	53549.99	0	53549.99	Resolved	Yes	Wealth Management - Incident Handler
INCIDENT-012	INCIDENT-012 - Title	No	Yes	Third Party	Third Party	No	Low	19479.34	211.97	19267.37	Under Review	No	Third Party - Incident Handler
INCIDENT-013	INCIDENT-013 - Title	No	No	Wealth Management	Wealth Management	Yes	High	2100000	97000	2003000	Resolved	No	Wealth Management - Incident Handler
INCIDENT-014	INCIDENT-014 - Title	No	No	Private Bank	Private Bank	No	Low	19.99	0	19.99	Resolved	Yes	Private Bank - Incident Handler
INCIDENT-015	INCIDENT-015 - Title	No	No	Corporate	Corporate	Yes	Medium	144000	4300	139700	Under Review	Yes	Corporate - Incident Handler
INCIDENT-016	INCIDENT-016 - Title	No	No	Premier	Premier	No	Low	9383.38	0	9383.38	Under Review	Yes	Premier - Incident Handler
INCIDENT-017	INCIDENT-017 - Title	Yes	No	Private Bank	Private Bank	No	Low	4239.39	0	4239.39	Under Review	No	Private Bank - Incident Handler
INCIDENT-018	INCIDENT-018 - Title	Yes	Yes	Third Party	Third Party	Yes	Low	26547.38	29.59	26517.79	In Progress	Yes	Third Party - Incident Handler
INCIDENT-019	INCIDENT-019 - Title	No	No	Private Bank	Private Bank	Yes	High	1336097.63	0	1336097.63	In Progress	Yes	Private Bank - Incident Handler
INCIDENT-020	INCIDENT-020 - Title	No	Yes	Third Party	Third Party	No	Low	13013.13	4484.39	8528.74	Resolved	No	Third Party - Incident Handler

The data displayed above was exported directly from an internal Barclays system that the organisation uses to track incidents and risk events throughout the bank. However, the data was then anonymised ahead of being included in the report to safeguard its confidentiality. The data allows all colleagues to track and view all past and present incidents and risk events within the organisation, to ensure that incidents and risk events are not constantly reoccurring – as this would highlight that there is a wider issue within the organisation. The columns display all the information that any user of the system may need, as explained below:

- Incident_ID – This is the unique ID that is given to each incident or risk event for users to distinguish each incident or risk event from one another. This field would often be referred to as the primary key since no two ID fields are ever the same which allows users to filter this field and find the exact incident or risk event that they are looking for.
- Incident_Title – This field summarises roughly what the incident or risk event is about, giving a one-sentence summary of the incident or risk event.
- Reoccuring_Event – This field is classed as a ‘Yes/No’ field meaning that when the user inputs the incident or risk event into the system, they only have the option to populate this field as either ‘Yes’ or ‘No’, allowing other users to identify if the incident or risk event has occurred before.
- Third_Party_Fault – Again this field is classed as a ‘Yes/No’ field. This allows users to see whether the incident or risk event occurred as a result of a third-party provider or at the fault of Barclays.
- Causal_Area – This field helps the user to understand the area of the organisation, or the third party, that is at fault for the incident or risk event occurring.
- Impacted_Area - This field helps the user to understand the area of the organisation, or the third party, that is impacted as a result of the incident or risk event occurring.
- Risk_Event – This is the ground truth label, again classified as a ‘Yes/No’ field. This was used for training the dataset as the objective of the project was to

identify which incidents within Barclays should remain as an incident and which should be classified as a risk event.

- Rating – The rating field is categorised as ‘Low’, ‘Medium’ or ‘High’. This allows users to understand the severity of the incident or risk event. The severity ratings are as follows:
    - Low – The incident or risk event occurred within one team and can quickly be resolved.
    - Medium – The incident or risk event occurred within a whole department and may take time to implement a solution.
    - High - The incident or risk event is a bank-wide issue.

This gives users an understanding of how the incident or risk event has been escalated.

- Initial_Loss_ - The initial loss field provides users with a view of the loss, if any, that the organisation faced as a result of the incident or risk event occurring.
- Recoveries – The recoveries field provides users with a view of any income that the organisation were able to retrieve, if any, because of the incident or risk event occurring.
- Net_Loss – The net loss field takes into consideration both the initial loss field and the recoveries field and provides the user with an overall view of the final financial impact if any, that the organisation faced because of the incident or risk event.
- Status – The status field is categorised as ‘Under Review’, ‘In Progress’ or ‘Resolved’. This allows the user to gain a view of which stage the incident or risk event is at and can give an indication of what has still to be done regarding dealing with the incident or risk event.
- Escalated_Appropriately - The field again is classified as a ‘Yes/No’ field, allowing the users to identify if the incident or risk event has been escalated up the chain and to the appropriate stakeholders. Each user understands the escalation ratings and depending on the rating field is dependent on how the incident or risk event must be escalated.
- Point_Of_Contact – The point of contact field allows the user to gain a view of whom they can contact if they have any questions, queries, or suggestions about the incident or risk event. This field displays a colleague’s name when pulled from

the system within Barclays, however, due to data confidentiality the name was anonymised and substituted with the individual's job role.

The risk event threshold is met when the 'Net_Loss' field displays a value that is equal to or greater than 25,000.00, all of the financial-related fields within the dataset are displayed as British Pound Stirling (GBP).

The dataset above was pulled from an internal system within the organisation that is used to record and manage all incidents and risk events against the organisation. However, the data has been anonymised since it is considered restricted and cannot be shared with individuals out with the organisation. Nevertheless, it is clear from the data presented that the column 'Risk_Event' clearly indicates whether the incident has met the criteria to be classified as a risk event. Having said that, this project was deemed effective to go underway since from the dataset outlined above there are lots of fields that many colleagues find to be irrelevant since the only piece of relevant information that colleagues are looking for is to identify whether the incident is classified as a risk event or not. This project became useful to carry out as the data displayed feeds into several forums and discussions that take place throughout the organisation, however, there is consistent feedback shared relating to the fact that the most crucial piece of information within the dataset is the 'Risk_Event' field and that the rest of the information is less essential to colleagues. Therefore, the dataset outlined above was used as part of a project that tells colleagues whether the incident has met the threshold for a risk event, without including all the additional information that can be found in the dataset above. However, before commencing with the project all the data outlined in the above dataset was assessed to ensure that it was accurate and up-to-date according to Jain (et al, 2020: p. 3561-3562) "the performance of a machine learning (ML) model is upper bounded by the quality of the data", emphasising the importance of assessing the data before the commencement of the project.

Data Pre-Processing

The data used as part of this can be exported, making it easier for the user to analyse multiple incidents at one time and compare incidents to ensure that the organisation does not have repeated incidents occurring.

However, all the data displayed throughout this assignment has been anonymised as data within Barclays PLC is labelled as 'Restricted – Internal' and therefore classified for staff within the organisation, 'Restricted – Internal' data cannot be shared with any individual outside of the organisation under any circumstances.

Ahead of the data being used as part of the project, it was first analysed to ensure that all fields were populated correctly to ensure the data was suitable enough to be used as part of the project. This was completed by filtering out incidents and risk events that occurred within the chosen time frame - the period between 1st January 2022 and 30th April 2022. Following this, the data was then reviewed by me (an analyst) to ensure that the data was of high quality – meaning that all the fields were filled out and completed as anticipated. There were no missing values as raising an incident or risk event means that the point of contact must complete and lengthy form with many mandatory fields requiring completion – making it difficult for the incident or risk event to be published with missing data. Once the analysis was completed the data was then saved down from an XLSX file to a CSV file before being uploaded into Jupyter. The data selected was a list of all the incidents that occurred in the first quarter of 2022, the period between 1st January 2022 and 30th April 2022.

Import Libraries

To begin with, the project, files, and models were implemented – as highlighted below:

```
In [182]: import pandas as pd
          df = pd.read_csv("AML Dataset.csv")
          df
```

Pandas were imported into the model since the model would be focusing on data tables and therefore pandas were the most applicable toolkit to implement.

```
In [176]: from sklearn.model_selection import train_test_split
```

Train-Test Split was used to divide the data into two subsets, one for training the model and another for testing it. Train-Test Split was used to validate the results as the project focuses on regression and classification.

```
In [179]: from sklearn.linear_model import LogisticRegression
```

Logistic Regression was implemented to get a view of a model that predicted the probability of the incident being classified as a risk event.

```
In [185]: from sklearn.preprocessing import LabelEncoder
```

Label Encoder was implemented into the Python code to convert labels into a numeric format, making it easier for the model to compute the output.

```
In [225]: from sklearn import tree
```

Following from importing Logistic Regression, a Decision Tree was implemented to confirm if the Logistic Regression model made the correct predictions. This is because a Decision Tree is more powerful than a Logistic Regression Classifier.

```
In [236]: import pandas as pd
          from matplotlib import pyplot as plt
```

Matplotlib was implemented as pyplot to create a plotting graph for the Scatter Plot to be displayed and indicate to users whether the incident is predicted to meet the decision boundary and separate the space of solutions that will be classified as ‘Yes’ (the incident meets the threshold of a risk event) or ‘No’ (the incident does not meet the threshold of a risk event).

Problem Type / Type of Problem – Predicting Incidents That Meet Risk Event Threshold

The goal of the project was to ensure that the model was easily able to identify if an incident met the threshold for a risk event. The threshold will mark if the solution is displayed in an area that represents risk events from others that are classed as incidents. This was a relevant project to carry out since one of the main

responsibilities of CCO is to identify incidents that qualify as risk events; therefore, it was relevant for a model to be able to tell the user if the incident met the risk event threshold – as well as an overall view of which incidents met the risk event threshold from a collection of set data.

This project was carried out to ensure the logistic regression model displayed the results that were intended. Geeks for Geeks define a regression problem as the output of the variable being “a real or continuous value, such as “salary” or “weight””. In this case, the value of £25K is used to identify whether an incident meets the threshold for a risk event within Barclays.

Why Regression Class Problem?

At the beginning of the project, it was decided that a logistic regression model would be implemented, as displayed below:

```
In [179]: from sklearn.linear_model import LogisticRegression

In [180]: model = LogisticRegression()

In [245]: model.fit(X_train,y_train)
Out[245]: DecisionTreeClassifier()

In [270]: model.predict(X_train)
Out[270]: array(['Yes'], dtype=object)
```

The python code above asks the machine to import a logistic regression model, the model that displays the probability of an event taking place – in this case, whether the incident meets the threshold for a risk event. The model is then asked to decide on whether or not the incident should be classified as a risk event by looking at both the ‘Incident_ID’ field and the ‘Net_Loss’ field, giving the user output as to whether or not the incident should be classified as a risk event. The model looks at both the ‘Incident_ID’ field and the ‘Net_Loss’ field to determine which specific incident meets the threshold, without including the ‘Incident_ID’ field the user would just be left with a result of ‘Yes’ or ‘No’ and would not be able to determine which specific incident meets the risk event threshold.

The above regression model was a useful model to implement as it allows an audience to view the outcome as to whether each incident meets the threshold for a

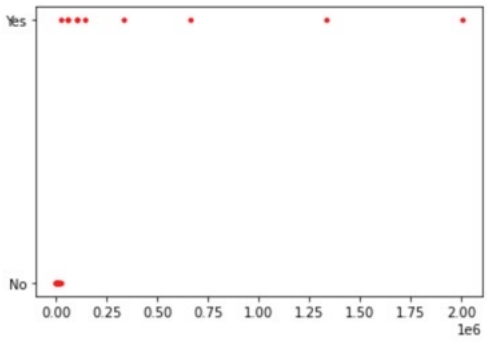
risk event, without already knowing the answer as to whether the incident is a risk event. This links with the scatter plot below, as we see several values spread across the x-axis (the axis running across the bottom of the graph) and up the y-axis (the axis running up and down the graph) we see ‘Yes’ and ‘No’. The red dots plotted throughout the graph indicate the results as to how many incidents potentially meet the threshold for a risk event and how many incidents do not, the red dots plotted nearer the top of the graph, in line with the word ‘Yes’ have taken into consideration the financial impact of the incident and estimated that this particular incident has met the threshold to be classified as a risk event – for instance, the incident has a financial value that is greater than or equal to £25K. However, the red dots plotted nearer the bottom of the graph, in line with the word ‘No’ have taken into consideration the financial impact of the incident and estimated that this particular incident has not met the threshold to be classified as a risk event – for instance, the incident has a financial impact that is less than £25K. It was decided that a logistic regression model is more applicable than any other model as it “describes the relationship between a dependent variable, y, and on or more independent variables, x” (MathWorks 2022). In this case, the dependent variable(s) (variable y) were the variables listed on the y-axis, in this instance ‘Yes’ and ‘No’, these variables are classed as dependent variables since variable x had to align to either ‘Yes’ or ‘No’. For example, the incident either did meet the £25K threshold and therefore should be classified as a risk event (‘Yes’) or it did not and therefore did not meet the threshold to be classified as a risk event (‘No’). However, the variable x in this case would be classed as the independent variable since the variable x could have been any number of things. In this case, the x variable had to be an amount and therefore made up of numbers. However, there was no limit on how large or small the number could be since incidents rarely have the same value as one another, making this variable the independent variable since there was no limit on variable x – the variable dotted across the x-axis. A logistic regression model was selected as the best machine learning approach as there are many advantages to a logistic regression model such as “(1) determining the strength of predictors, (2) forecasting an effect, and (3) trend forecasting” (Statistic Solutions, No Date). These advantages define the purpose of a logistic regression model, this model was selected to gain a view of which incidents were predicted to be classed as a risk event.

Scatter Plot

The main goal of the project was to determine whether each incident within Barclays PLC met the threshold to be classified as a risk event. Therefore, it became applicable to implement a scatter plot into the project. The scatter plot became a useful plot to implement since it was used to support the machine in learning by allowing the model to predict how many incidents would meet the risk event threshold and how many would not. Although not going into detail regarding which incidents did meet the threshold for a risk event and which did not, the scatter plot displays an overall view of the number of incidents that meet the risk event threshold and the number of incidents that do not. Scatter plots are useful tools to implement as they show "how much one variable is affected by another or the relationship between them with the help of dots in two dimensions" (Tutorials Point, No Date). This was implemented as it identifies the relationship between the variables and the clear results that a scatter plot can highlight. In addition to this, Cliche (et al, 2017: p. 135-150) explained that "charts are an excellent way to convey patterns and trends in data", thus supporting the idea that a scatter plot is a sensible chart to include within the model as it displays any themes and/or trends from the data without having to read through a large amount of data.

A scatter plot was implemented into this particular project to get a closed in view of the 'Net_Loss' field and the 'Risk_Event' field, these two fields, in particular, were brought into focus since the 'Net_Loss' field is ultimately the field that tells the analyst whether the incident has met the threshold for a risk event – a reminder that an incident is classed as a risk event when the 'Net_Loss' field displays a value that is greater than or equal to £25K. The scatter plot was used to test the machine and to get a view of whether the machine was able to predict which incidents met the threshold to become a risk event and which did not, as displayed below:

```
In [242]: plt.scatter(df.Net_Loss,df.Risk_Event,marker='.',color='red')
Out[242]: <matplotlib.collections.PathCollection at 0x7fbeae23eb50>
```



The scatter plot above shows an overall prediction of which incidents meet the financial threshold to be deemed as a risk event, compared to the incidents that do not meet the financial threshold to be classed as a risk event. These predictions tell the user that of the nineteen incidents recorded, eight of them meet the financial threshold to be classified as a risk event – without telling the user which of the eight incidents meets the threshold.

Following this, the data was then broken down to display each incident and whether the incident is a risk event or not. The broken-down data can be found highlighted in the below graph, along with the python code used to display the data.

```
In [183]: inputs = df.drop('Risk_Event',axis='columns')
          target = df['Risk_Event']

In [184]: target
Out[184]: 0      No
          1      No
          2      No
          3      No
          4     Yes
          5     Yes
          6      No
          7     Yes
          8     Yes
          9     Yes
          10    Yes
          11     No
          12    Yes
          13     No
          14    Yes
          15     No
          16     No
          17    Yes
          18    Yes
          19     No
          Name: Risk_Event, dtype: object
```

The scatter plot previously inserted gives a view of the predicted probability of the incident being classified as a risk event, however, the scatter plot only provides the user with an overall view of the overall number of incidents that meet the threshold to

be classified as a risk event and the number of incidents that do not meet the threshold to be classified as a risk event. From the scatter plot the user can't identify which incidents do meet the risk event threshold and which do not, therefore the above was implemented to show the user exactly which incidents do meet the threshold to be classified as a risk event and which do not. The above was also implemented to allow the user to conduct a manual check back to ensure that the same number of incidents were predicted to be classed as risk events as displayed in the scatter plot - this was also useful to have implemented during the testing stage. The data was displayed in the table above by extracting the 'Risk_Event' field, which already determines whether the incident meets the risk event threshold. However, extracting this meant that the correct results were pulled through to the model and that a manual checkback could be conducted, meaning that when the programme was fully implemented it could be identified that the model was doing its job how it was intended by cross-referencing the result from the decision tree with the risk event column in the dataset.

Numerical Encoding

Throughout the project the data was scaled using a minimum and maximum value, this was carried out to support the data to better fit into the model. The minimum value used was 0 and the maximum value used was 19, the range of 0-19 was selected to ensure that each unique incident was still identifiable, and a unique primary key continued to exist within the data.

The below three screenshots display the Python code that was written following the implementation of the label encoder into the programme. The label encoder was implemented to convert the data into a numerical format, making it easier for the machine to read and translate - the below code shows the fields that the label encoder was aligned to and asked to translate into a numerical format. The machine was asked to convert the full dataset into a numerical format so that the user would have a complete view of the dataset translated into a numerical format, the machine was requested to do this so that when the decision tree was implemented at the end of the project the user would only have to insert the numerical value of each field

rather than having to insert the alphanumeric value of each cell in the dataset – which would take significantly longer for the user.

```
In [224]: inputs_ = inputs.drop(['Incident_ID','Incident_Title','Reoccuring_Event','Third_Party_Fault','Causal_Area','Impacted
          inputs_

In [224]: 'Causal_Area','Impacted_Area','Rating','Initial_Loss_','Recoveries','Net_Loss','Status','Escalated_Appropriately','P

In [224]: ting','Initial_Loss_','Recoveries','Net_Loss','Status','Escalated_Appropriately','Point_Of_Contact'],axis='columns')
```

The following two screenshots display the results of the label encoder being implemented on the dataset, meaning that the dataset is now displayed in a numerical format.

Out[224]:

	IncidentID_	IncidentTitle_	Incident_ID_	Incident_Title_	Reoccuring_Event_	Third_Party_Fault_	CausalArea_	ImpactedArea_	Rating_	Causal_Area_	Impac
0	0	0	0	0	0	0	5	5	1	5	
1	1	1	1	1	0	1	4	4	1	4	
2	2	2	2	2	0	0	3	3	1	3	
3	3	3	3	3	0	0	3	3	1	3	
4	4	4	4	4	0	1	4	4	0	4	
5	5	5	5	5	0	1	4	4	1	4	
6	6	6	6	6	0	0	1	1	1	1	
7	7	7	7	7	1	0	0	0	2	0	
8	8	8	8	8	0	0	0	0	2	0	
9	9	9	9	9	0	0	2	2	2	2	
10	10	10	10	10	0	0	5	5	1	5	
11	11	11	11	11	0	1	4	4	1	4	
12	12	12	12	12	0	0	5	5	0	5	
13	13	13	13	13	0	0	3	3	1	3	
14	14	14	14	14	0	0	1	1	2	1	
15	15	15	15	15	0	0	2	2	1	2	
16	16	16	16	16	1	0	3	3	1	3	
17	17	17	17	17	1	1	4	4	1	4	
18	18	18	18	18	0	0	3	3	0	3	
19	19	19	19	19	0	1	4	4	1	4	

Out[224]:

usalArea_	ImpactedArea_	Rating_	Causal_Area_	Impacted_Area_	Recoveries_	Net_Loss_	Status_	Escalated_Appropriately_	Point_Of_Contact_	Initial_Loss_
5	5	1	5	5	8	9	1	1	5	10
4	4	1	4	4	0	2	0	1	4	2
3	3	1	3	3	9	3	0	1	3	8
3	3	1	3	3	5	5	0	1	3	5
4	4	0	4	4	11	17	0	1	4	17
4	4	1	4	4	10	12	2	1	4	12
1	1	1	1	1	1	0	2	1	1	0
0	0	2	0	0	0	16	2	0	0	15
0	0	2	0	0	13	14	1	1	0	16
2	2	2	2	2	4	13	1	1	2	13
5	5	1	5	5	0	11	1	1	5	11
4	4	1	4	4	3	8	2	0	4	7
5	5	0	5	5	12	19	1	0	5	19
3	3	1	3	3	0	1	1	1	3	1
1	1	2	1	1	6	15	2	1	1	14
2	2	1	2	2	0	7	2	1	2	4
3	3	1	3	3	0	4	2	0	3	3
4	4	1	4	4	2	10	0	1	4	9
3	3	0	3	3	0	18	0	1	3	18
4	4	1	4	4	7	6	1	0	4	6

Data - Training and Testing

After successfully converting all the data into a numerical format, it then became important to test the machine. Testing the machine is a crucial part of the process since testing the machine “gives confidence, meaning organisations can progress with projects, lift productivity and help drive long-term success”, (IT Pro, 2021). Not only this, but it was important to test the machine to see if the machine could pick out the key pieces of information from the dataset to display to the user whether the incident met the threshold of a risk event, as set out by Barclays.

Decision trees were implemented in the machine to support part of the final output of the project. “Decision trees are a type of supervised machine learning” (Xoriant, 2019), this means that you give the machine the input and options for an appropriate output and the machine will decide based on the variables that are input. Decision trees can sometimes be complex and give multiple decisions and outputs. However, in this case, the decision tree only had one decision to make, the model was required to choose as to whether the incident met the threshold to be classified as a risk event. In simpler terms the machine was asked to give a ‘Yes/No’ type of output, ‘Yes’ being that the incident meets the threshold to be classed as a risk event and ‘No’ being that it doesn’t. The decision tree was displayed simply as its main purpose was to allow the model to give a ‘Yes/No’ type output. However, as well as this

Navada (et al, 2011: p.37-42) found that with decision trees, “as the complications increase its accuracy to make good Decision trees decreases”, highlighting that simpler decision trees are more accurate and effective than complex decision trees. Not only this but the decision tree was also set up simply to ensure that colleagues who are less familiar with machine models were able to make use of the model, the code highlighted below demonstrates the simple nature of inserting the value of the columns within the same row and receiving the output in the form of a ‘Yes/No’ answer – making it easier to understand for individuals who are not familiar with decision tree models.

```
In [225]: from sklearn import tree

In [226]: model = tree.DecisionTreeClassifier()

In [230]: model.fit(inputs_,target)
Out[230]: DecisionTreeClassifier()

In [231]: model.score(inputs_,target)
Out[231]: 1.0

In [232]: model.predict([[19,19,19,19,0,1,4,4,1,4,4,7,6,1,0,4,6]])
Out[232]: array(['No'], dtype=object)
```

The python code displayed below shows how the decision tree was implemented and how the programme was tested. Testing was carried out on incident number 19, which displays the result that this incident does not meet the threshold to be classed as a risk event, from the dataset below we know that this is true as we see that incident 19 does not meet the threshold to be classified as a risk event. In [232] from the above is the line in which the user inserts the values. With incident number 19 being the tested incident, we can see from the above that to test the decision tree all the numbers from row 19 were inserted into In [232] and that the results displayed Out [232] show that this incident does not meet the threshold to be classified as a risk event and should therefore remain categorised as an incident.

Out[224]:

	IncidentID_	IncidentTitle_	Incident_ID_	Incident_Title_	Reoccuring_Event_	Third_Party_Fault_	CausalArea_	ImpactedArea_	Rating_	Causal_Area_	Impac
0	0	0	0	0	0	0	5	5	1	5	
1	1	1	1	1	0	1	4	4	1	4	
2	2	2	2	2	0	0	3	3	1	3	
3	3	3	3	3	0	0	3	3	1	3	
4	4	4	4	4	0	1	4	4	0	4	
5	5	5	5	5	0	1	4	4	1	4	
6	6	6	6	6	0	0	1	1	1	1	
7	7	7	7	7	1	0	0	0	2	0	
8	8	8	8	8	0	0	0	0	2	0	
9	9	9	9	9	0	0	2	2	2	2	
10	10	10	10	10	0	0	5	5	1	5	
11	11	11	11	11	0	1	4	4	1	4	
12	12	12	12	12	0	0	5	5	0	5	
13	13	13	13	13	0	0	3	3	1	3	
14	14	14	14	14	0	0	1	1	2	1	
15	15	15	15	15	0	0	2	2	1	2	
16	16	16	16	16	1	0	3	3	1	3	
17	17	17	17	17	1	1	4	4	1	4	
18	18	18	18	18	0	0	3	3	0	3	
19	19	19	19	19	0	1	4	4	1	4	

Out[224]:

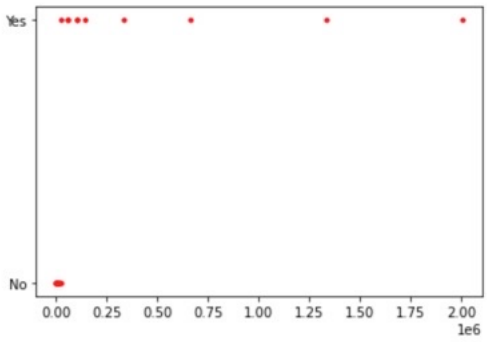
usalArea_	ImpactedArea_	Rating_	Causal_Area_	Impacted_Area_	Recoveries_	Net_Loss_	Status_	Escalated_Appropriately_	Point_Of_Contact_	Initial_Loss__
5	5	1	5	5	8	9	1	1	5	10
4	4	1	4	4	0	2	0	1	4	2
3	3	1	3	3	9	3	0	1	3	8
3	3	1	3	3	5	5	0	1	3	5
4	4	0	4	4	11	17	0	1	4	17
4	4	1	4	4	10	12	2	1	4	12
1	1	1	1	1	1	0	2	1	1	0
0	0	2	0	0	0	16	2	0	0	15
0	0	2	0	0	13	14	1	1	0	16
2	2	2	2	2	4	13	1	1	2	13
5	5	1	5	5	0	11	1	1	5	11
4	4	1	4	4	3	8	2	0	4	7
5	5	0	5	5	12	19	1	0	5	19
3	3	1	3	3	0	1	1	1	3	1
1	1	2	1	1	6	15	2	1	1	14
2	2	1	2	2	0	7	2	1	2	4
3	3	1	3	3	0	4	2	0	3	3
4	4	1	4	4	2	10	0	1	4	9
3	3	0	3	3	0	18	0	1	3	18
4	4	1	4	4	7	6	1	0	4	6

Results

The results of the project were positive as the models implemented in this project link together to provide Barclays with a useful tool that can be used to analyse all incidents as a whole, through the implementation of the scatter plot. The scatter plot shows an overall view of all of the incidents from the data cut and tells the user whether the incident qualifies as a risk event, the informed decision is made by the model by analysing both the ‘Net_Loss’ field and the ‘Risk_Event’ field from the data cut. The data cut from the CSV file has many fields that become unnecessary if the user is only using the data to determine if each incident qualifies as a risk event, hence why the decision was made to have an overall view of all the incidents that do not meet the threshold for a risk event and to display the incidents that are classed

as a risk event. However, although the scatter plot is a useful tool to implement to allow the user to have an overall view of the incidents and risk events, the downside of scatter plots is that the overall view does not indicate as to which incidents do or do not meet the risk event threshold.

```
In [242]: plt.scatter(df.Net_Loss,df.Risk_Event,marker='.',color='red')
Out[242]: <matplotlib.collections.PathCollection at 0x7fbeae23eb50>
```



Linking back to the scatter plot where it was highlighted that the graph only gives an overall view of the number of incidents that do or do not meet the criteria to be classed as a risk event, compared to a breakdown of each incident or risk event. Hence why the table below was implemented into the project which gives the user a view of which incidents do or do not meet the risk event threshold:

```
In [183]: inputs = df.drop('Risk_Event',axis='columns')
          target = df['Risk_Event']

In [184]: target
Out[184]: 0      No
          1      No
          2      No
          3      No
          4     Yes
          5     Yes
          6      No
          7     Yes
          8     Yes
          9     Yes
          10    Yes
          11     No
          12    Yes
          13     No
          14    Yes
          15     No
          16     No
          17    Yes
          18    Yes
          19     No
          Name: Risk_Event, dtype: object
```

This was implemented so that the user can get a clear view of the status of each incident/risk event without having to jump between the model and the data cut to determine the status of each incident or risk event.

Following this, the CSV data cut was scaled using a minimum and maximum value to help support the machine when implementing a decision tree, which followed the above step. The decision tree allowed the user to implement the values of each of the variables and receive an output telling the user whether the incident has met the criteria of a risk event. The decision tree indicates this to the user by providing an answer of 'Yes', the incident has met the criteria of a risk event, or 'No', the incident has not met the criteria of a risk event and will remain classified as an incident.

There is strong evidence from the data testing and training that was conducted that the implemented model is accurate with no error rate, this is to the fact that each incident was tested in the model and each of the results came out as expected – giving the model a 100% accuracy rating. However, there were errors within the python code, to begin with, which as a result forced the decision tree to provide false results. Having said that, these errors were quickly resolved, allowing the model to be complete and testing to begin, which showed a 100% success rate, as previously stated.

Limitations

Despite the outcome of the project being a positive one, there were limitations to this project. Firstly, the sample size of data that was used was fairly small as it only consisted of twenty incidents. Although this may not sound small there are thousands of incidents logged on the Barclays system, so considering this the sample size of the project would be considered fairly small. It was previously stated that after the testing and training of the data it was found that the machine was 100% accurate at identifying incidents that meet the risk event threshold. However, had the sample size been larger than it was there would have been a higher chance that the model may not have had this level of accuracy, along with this a larger sample size would mean that not all incidents could have been individually tested.

A second limitation to this project is that the CSV data cut that was used as part of this project already has a field that tells the user whether or not the incident has met the threshold of a risk event, the objective of this project was to identify if the

machine was able the detect if an incident met the threshold of a risk event to ensure that the results matched up. However, if the 'Risk_Event' field was not included in the data cut then it could have made it harder for the model to identify the status of the incident/risk event. Having said this, the necessary steps were taken at the key stages of the project to ensure that this field was ignored to determine if the model was able to identify for itself whether or not the incident met the criteria of a risk event.

Conclusion

Organisations across the world of all different sizes, successes, and across all sectors face the threat of incidents or risk events occurring within the organisation. Incidents and risk events can not only result in reputational and/or financial impact on the organisation but can also have severe consequences on an organisation's customers, clients, and colleagues and it is vitally important that all organisations manage incidents and risk events professionally and appropriately. This is particularly important in financial institutions such as Barclays because financial institutions are heavily regulated and continually monitored by organisations such as the Financial Conduct Authority (FCA). Throughout this project, many models were implemented using Python programming to sample a cut of incidents and risk events that Barclays have faced. Techniques such as scatter plots, logistic regression and decision trees were implemented to create a model that can identify which incidents meet the threshold to be classified as a risk event – by ensuring that the machine understood that a risk event occurs when there is a net loss greater than or equal to £25K. This machine learning project was found to be 100% effective, as proven through training, and testing of the model with the addition that this program was found easy to use and understandable by colleagues who do not usually work with machine learning programs or code with Python.

References

Barclays

No Date

Barclays Office Locations

https://home.barclays/careers/barclays-office-locations/

[Date Last Accessed: 4th May 2022]

Barclays

No Date

Who We Are and Our Values

https://www.banking.barclaysus.com/who-we-are.html

[Date Last Accessed: 4th May 2022]

Cliche, M., & Rosenberg, D., & Madeka, D., & Yee, C. (2017)

Scatteract: Automated Extraction of Data from Scatter Plots

Joint European Conference on Machine Learning and Knowledge Discovery in Databases

Pages 135-150

https://link.springer.com/chapter/10.1007/978-3-319-71249-9_9

[Date Last Accessed: 17th May 2022]

Geeks for Geeks

No Date

Regression and Classification | Supervised Machine Learning

https://www.geeksforgeeks.org/regression-classification-supervised-machine-learning/amp/

[Date Last Accessed: 8th May 2022]

IT Pro

2021

The Importance of Data Testing and How To Do It Right

https://www.itpro.co.uk/business-strategy/data-insights/358811/the-importance-of-data-testing-and-how-to-do-it-right

[Date Last Accessed: 21st May 2022]

Jain, A., & Patel, H., & Nagalapatti, L., & Gupta, N., & Mehta, S., & Guttula, S., & Mujumdar, S., & Afzal, S., & Mittal, R., & Munigala, V. (2020)
Overview and Importance of Data Quality for Machine Learning Tasks
Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining
Pages 3561-3562
https://dl.acm.org/doi/abs/10.1145/3394486.3406477
[Date Last Accessed: 6th May 2022]

MathWorks
2022
What Is A Linear Regression Model?
https://www.mathworks.com/help/stats/what-is-linear-regression.html
[Date Last Accessed: 10th May 2022]

Navada, A., & Ansari, A., & Patil, S., & Sonkamble, B. (2011)
Overview of Use of Decision Tree Algorithms in Machine Learning
2011 IEEE Control and System Graduate Research Colloquium
Pages 37-42
https://ieeexplore.ieee.org/abstract/document/5991826
[Date Last Accessed: 30th May 2022]

ScrapeHero
2022
Barclays Bank Locations in the UK
https://www.scrapehero.com/store/product/barclays-bank-locations-in-the-uk/
[Date Last Accessed: 8th May 2022]

Statista
2022
Barclays Bank – Statistics & Facts
https://www.statista.com/topics/3913/barclays-bank/#topicHeader_wrapper
[Date Last Accessed: 14th May 2022]

Statistics Solutions
No Date
What is Linear Regression?
https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/what-is-linear-regression/
[Date Last Accessed: 11th May 2022]

Tutorials Point
No Date
Machine Learning – Scatter Matrix Plot
https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_scatter_matrix_plot.htm
[Date Last Accessed: Date Last Accessed: 17th May 2022]

Xoriant
2019
Decision Trees for Classification: A Machine Learning Algorithm
https://www.xoriant.com/blog/product-engineering/decision-trees-machine-learning-algorithm.html
[Date Last Accessed: 30th May 2022]