School of Computing
FACULTY OF ENGINEERING

UNIVERSITY OF LEEDS

Exploratory data analysis and Regression model for Navigation dataset

Arjun Manjunatha

Submitted in accordance with the requirements for the degree of
MSc Advanced Computer Science

2021/2022

The candidate confirms that the following have been submitted*:*

Items	**Format**	**Recipient(s) and Date**
Deliverables 1	Report	SSO (19/08/2022)
Deliverables 2	Jupyter notebook for exploratory data analysis URL: https://github.com/arjunm97/Thesis-Project-University-of-Leeds	Supervisor, assessor (19/08/2022)
Deliverables 3	Jupyter notebook for Regression models URL: https://github.com/arjunm97/Thesis-Project-University-of-Leeds	Supervisor, assessor (19/08/2022)

Type of Project: Empirical Investigation

The candidate confirms that the work submitted is their own and the appropriate credit has been given where reference has been made to the work of others.

I understand that failure to attribute material which is obtained from another source may be considered as plagiarism.

(Signature of student) Arjun Manjunatha

© 2022 The University of Leeds and Arjun Manjunatha

Summary

Dementia is the decline of brain functioning which affects the quality of life for a lot of people. It is usually prevalent in the older population. The first sign of dementia is the loss of spatial awareness. A study was conducted by UCL to test the navigational ability of players, this was done to quantify the navigational ability and come up with a method to detect the onset of dementia by analyzing the navigational ability of the subject. They built a game called Sea hero Quest to achieve this. The game has several levels, each level testing the navigational ability. They gathered data on the navigational ability of players from different backgrounds and demographics. They also gathered insights into the navigational ability of different players. They were not able to build a model to detect the onset of dementia but they set the stage for it. In this project, the main goal is to bridge the gap between the final model to detect the onset of dementia and the work done by the UCL study.

In this project, exploratory data analysis is conducted to gain insights into the navigational dataset by understanding the relationships between different data points. For instance, the project tries to find the correlation between age and performance in the game. These insights will help in understanding the pattern that can be used to build a model to predict the onset of dementia. Also in this project, regression models are built to understand to predict the final score of the player from the player's initial level scores. There were also experiments conducted to classify the performance of the player using Naive Bayes, Random Forest, and Decision Tree.

Finally, the results are discussed from the exploratory data analysis stage. Insights such as the performance in the city hero quest and sea hero quest for the younger population are similar but for the older population, the results are not similar. Also how education level affects the performance in the game for males and females. The regression models built also perform with low error metric values, that is the regression model's predictions are very close to the actual values. The results are compared for different models and the reasons for the results are also discussed.

Acknowledgments

I would like to thank my supervisor, Professor Tony Cohn who has been very helpful and supportive thorough out the project. He gave me helpful feedback periodically. His guidance helped me to approach the project in a systematic and structured manner. I would also like to thank my assessor, Dr. Yanlong Huang, for giving me useful feedback and inputs.

I would like to thank my parents Dr. Majunatha and Dr. Revathi for supporting me throughout my life and believing in me. It would have not been possible to do my masters without the support of my parents.

I would also like to thank the computer science department at the University of Leeds for being helpful and supportive in my studies during my postgraduate year.

Table of Contents

List of Figures

List of Tables

Abbreviations: SHQ- Sea Hero Quest Score, CHQ- City Hero Quest Score

Chapter 1 Introduction

1.1 Context

A study was initially conducted by UCL to predict the initial signs of Dementia. This was done by creating a game called Sea Hero Quest(Coutrot et al., 2018). According to the study, the first sign of Dementia is loss of spacial orientation(Coutrot et al., 2018). Hence they created a game called Sea Hero Quest, In this game, the players have to navigate through different islands, and the app tests the navigation ability of different players from 80 levels. The study then collected demographic and performance metrics of the players. Then they analyzed different age group's performance and the performance between males and females, they also studied how education level affects performance.

The paper “Entropy of city street networks linked to future spatial navigation ability”(Coutrot et al., 2022) conducted further research on the topic above. They analyzed how a player’s environment plays a role in the performance of the game. They also introduced a new game where city streets are used instead of sea routes to test the player’s navigational ability. This game was called City Hero Quest(Coutrot et al., 2022).

This project will analyze the data produced by the above studies. That is find interesting insights of the sea hero quest and city hero quest datasets using data science techniques. This project also uses machine learning to predict the final city hero score by using the initial level scores.

1.2 Motivation

Technology is a powerful tool in alleviating human suffering in many ways. Dementia is recognized as a global public health problem and does not have a definitive cure as of now. Early diagnosis of the condition may help preserve the quality of life. The study conducted by UCL(Coutrot et al., 2018) is a pioneering one, which has thrown insights into the possibilities of developing an early detection model by testing a person's navigational abilities.

Finding more meaningful insights and patterns from the player's navigational abilities will help in building models which can detect the early onset of dementia. Although the main goal of the study conducted by UCL(Coutrot et al., 2018) was to detect the early onset of dementia, there is a lack of data points to build that model. But taking small steps such as

understanding the relationship between the performance and different factors will give insights that can be useful in building models in the future. With the help of insights and preliminary models explored, in the future, there can be a model that predicts the onset of dementia by analyzing the performance in the Sea hero quest game. This project hopes to bridge the gap between the ultimate goal of detecting the early onset of dementia and the current stage of research.

1.3 Aim

There are two main goals for this project.

1)The first goal is to use exploratory data analysis to find meaningful insights from City hero quest dataset.

2)The second goal is to build a machine learning model to predict the final City Hero Quest score from the initial levels.

In order to achieve these goals, Python programming language is used. Linear regression and its variations are used for the prediction of the final City Hero Quest score.

1.4 Objectives

1) Conduct literature survey and get information on what could be done and what is already achieved.

2) Gather the dataset and familiarise with the dataset.

3) Data preprocessing and feature engineering.

4) Conduct Exploratory data analysis to find insights.

5) Build machine learning models for prediction.

6) Evaluate the results of the model and the insights gathered from the exploratory data analysis stage.

7) Deliver the report and code.

1.5 Deliverables

1) Jupyter notebook containing the code for exploratory data analysis.

2) Jupyter notebook containing the code for prediction models.

3) A report. The report contains information on the previous research conducted related to the aim of the project, what techniques were used, how the experiments were set up, and finally, the results obtained.

1.6 Report Structure

1) Chapter 1 gives a context of the project. It also provides the aim and objectives of the project. Lastly includes deliverables at the end of the project.

2) Chapter 2 briefly specifies the background information. Covers the techniques and tools available to implement the project. The last section provides information on which techniques were used in the project.

3) Chapter 3 explains how the datasets were preprocessed. How the dataset was used to conduct exploratory data analysis. Also, explains the steps to build the models which predict the CHQ scores.

4) Chapter 4 discusses the insights obtained from the data exploratory analysis stage. And also the results of the prediction model.

5) Chapter 5 is about how the results obtained were verified and how it was validated.

6) Chapter 6 gives a conclusion for the project and suggests what future work could be done.

1.7 Relevance to the degree

With the help of two modules studied namely Programming for data science(COMP5712M), and Machine Learning (COMP5611M) this project was completed. In the module, programming for data science, the techniques learned for data exploration and programming

helped. While from the module Machine learning, the concepts of building a machine learning model were used.

1.8 Ethical, Legal and Social Issues

There are no ethical, legal, or social issues with this project

Chapter 2 Background

2.1 Literature survey

The first paper referred to was Global Determinants of Navigation Ability (Coutrot et al., 2018), researchers developed a game to test navigational ability which involved analyzing the spacial environment. Subjects were picked from various countries and from different cultural backgrounds. The performance in the game was used to find five distinct correlations; those were, the wealth of the country to their performance in the game, gender's performance in the game, and lastly age groups and their performance in the game. The researchers used a metric called "overall performance corrected" (OPcorr)(Coutrot et al., 2018) to assign a number to the navigational ability or performance of the player. The game Sea Hero Quest was created to test the spatial navigation ability of the players. The game was used to measure the performance in two main tasks, path integration, and wayfinding. There was a set path in wayfinding with checkpoints where the players had to memorize the way. In the other task, the players had to follow the flair gun and in the end, had to return to the starting point. The path used and the demographics of the players were collected along with a few other performance metrics. They used a clustering algorithm to cluster different groups(Coutrot et al., 2018).

They used the same game to test if they would be able to predict the onset of dementia(Coutrot et al., 2018). As they said the first sign of dementia was the loss of spacial orientation. They wanted to measure the performance of the players and predict if the subject would be in the first stage of dementia. They analyzed the data and gathered insights such as the performance comparison between genders, education level, and their performance, finally measuring the performance among different geographies.

In this paper "Entropy of city street networks linked to future spatial navigation ability"(Coutrot et al., 2022), the researchers explored the correlation between geographic location and the performance in navigation ability task which was again emulated by the game Sea Hero Quest. They also created another game called City hero Quest, where instead of sea they emulated the streets of a city to conduct their experiment. They concluded that people who are born in rural areas are better at navigating rural areas while people who grew up in urban areas are better at urban navigation(Coutrot et al., 2018). They assigned a number for each city based on several factors such as type of layout of the city, number of unique streets, betweenness, closeness and degree centrality, average circuity, average

neighborhood degree, clustering coefficient, and average street length, and several others. The number which they assigned was called the entropy of the city. They collected data on the performance of City Hero Quest alongside Sea Hero Quest for 600 players. This dataset is used for the project.

2.2 Methods and Techniques

There are many methods, techniques, and tools available for achieving the goals of this project. Which will be discussed in this section. For the entire project Python programming language is used. For exploratory data analysis libraries such as Numpy, Pandas, Seaborn, Plotly, and Matplolib can be used. Numpy library helps with the processing of arrays and also statistical insights of the array can be easily found using inbuilt functions, for instance, to find the mean of the array there is an inbuilt function that comes with the library. Seaborn, Matplotlib, and Plotly are libraries used to plot graphs and visually represent the data. Some data can be better interpreted visually and this library comes in handy for handling those cases. The library can be used to plot various types of graphs such as scatter plots, line graphs, bar charts, and many more. Pandas library is used to create a data frame which is a table-like structure of the dataset. Pandas is used because of its versatility. Slicing of dataset or selecting specific rows based on conditions can be easily done in a pandas data frame.

For the data preprocessing part, there are again many tools and techniques available. The first step is to load the dataset. This can be achieved by reading the CSV file or connecting to an API or connecting to a server, all this can be handled by python and its libraries. The dataset might have missing values which are taken care of by finding the mean or median of the column and replacing it in place of missing values. Datasets might have categorical values for columns, for example, a column “city” might have names of cities as values, which needs to be encoded before using the data for training machine learning models. The dataset needs to be split into training and testing before its feed to the machine learning algorithm. Also, the outliers need to be found and removed for visualization and for training models. In some datasets, particular columns might have very high values compared to the other columns, for example, salary and age columns might have very different scales, before the dataset can be used it needs to be scaled down, this is called normalization. The last step is feature engineering, in this step techniques such as visualization and clustering can be used to determine which features can be used for training the model.

Machine learning algorithms can be used to create models to predict values or classify them into categories by feeding the models with data and training them. There are two types of

outcomes; prediction, and classification. There are three types of machine learning algorithms supervised, unsupervised and semi-supervised. In supervised learning, the algorithm is told what to predict, in the training stage there is a pre-determined input-output pair. While in unsupervised learning the model is not given predetermined information. And finally, in semi-supervised learning, there is a small amount of labeled data and unlabelled data. There are mainly ten types of regression algorithms and five types of classification algorithms.

Linear regression, Lasso regression, Ridge regression, Gaussian regression, Polynomial regression, artificial neural network, and many more can be used to predict continuous values. In regression, the line of best fit for all the points in the dataset is used to create an equation. When predicting, the values are substituted into the equation to get the expected value.

Naive Bayes calculates the probability of whether a data point belongs to a certain category or not. KNN is a pattern recognition algorithm that finds the nearest neighbors of a data point and all the points nearest to it become a class. The decision tree creates nodes and leafs to classify the data into different categories. The leafs of a decision tree are the final output classification.

Artificial neural networks can also be used to predict and classify data. In artificial neural networks, there are input, output, and hidden layers. The hidden layers learn the information provided by the input layer by weighting the information and by also using the backpropagation technique. When the network is trained and it is able to predict values from test data.

Finally, some of the techniques and methods which can be used to validate the results are performance metrics such as accuracy, precision, recall, confusion matrix, least square methods, Dunn’s index, Silhouette coefficient, Adjusted R square, Root mean square error, Mean absolute error, Precision recall curve, and Logarithmic loss. These are some of the many more performance metrics. Accuracy is the number of true positives and true negatives divided by the number of true positives, true negatives, false positives, and false negatives. Precision is the number of positive class predictions that actually belong to the positive class. Mean Absolute Error is the average of the difference between the original value and the predicted value. All the metrics available are not discussed in this section as there are many of them and the ones used are defined and discussed on why they were used in the next section.

2.3 Choice of Methods

All the code and scripts are written in python because there are libraries available to handle data, build models, and for visual representations. Pandas library is used for this project because of the ease with which datasets can be loaded and the ease of conducting data analysis. For instance, slicing the datasets or considering specific columns for certain analyses can be easily done in pandas. Matplotlib and seaborn were used for data visualization. Specifically scatter plots and bar graphs options were used from these libraries for data visualization. Seaborn was used because it had the zooming and customizing the plot option which is helpful in finding finer insights from the graph.

Sklearn library was used to build all the machine learning models in this project. It was used because it has all the complex code for the algorithm pre-written. The user has to import the class and initiate it and train the model. Also, it provides the option to fine-tune the model to get better results.

Multivariate regression, lasso, and ridge regression were used in this project to build prediction models. The model takes in the previous level scores and predicts the final score based on the previous levels. All the values are continuous in nature and also have decimals. Regression is built to handle continuous values while a classification algorithm is not good at handling continuous values(Li et al., 2022). Even though the clustering algorithm seems like a good option looking at the dataset. Clustering algorithms are not good at handling continuous values(Li et al., 2022).

Regression is a method for understanding the relationship between independent variables or features and a dependent variable or outcome. Outcomes can then be predicted once the relationship between independent and dependent variables has been estimated(Machine Learning Regression Explained - Seldon, 2022). Lasso and ridge are variations of linear regression, but in these regression models, there is a penalty for independent variables which has a low impact on the target variable. Also, Lasso and ridge are used for solving the problem of overfitting(Lasso and Ridge Regularization - A Rescuer From Overfitting, 2022).

There are also some experiments in this project where the data is classified into “good” and “fair” performances. Hence to categorize the performance, Random Forest and Naive Bayes algorithms are used. An Artificial Neural network was not used to predict because the dataset is very small with 600 players’ data, which can be considered a small dataset for an artificial neural network, while machine learning works relatively well for smaller datasets. All

the performances of both regression and classification algorithms are compared and later discussed in the results chapter.

The performance metrics used in this project are mean absolute error, root mean error squared, and mean squared error. These metrics are used for regression models. These metrics paint a good picture of how well the regression models are performing. The error metrics are used for regression because it suggests how far off the predictions are from the actual values. Since there is no absolute method to quantify the results like that of classification model metrics, error metrics are used. Similarly, for classification algorithms, accuracy, precision, recall, and confusion metrics were used, as these measures give a good picture of how well the model is performing.

Chapter 3 Datasets and Experimental Design

3.1 Overview

The first section is about the dataset, its variables, and information about the dataset. Further, the chapter discusses how exploratory data analysis was conducted. Following this, feature engineering for the regression model is discussed. The next few sections explain how the multivariate regression and regularized regressions are built to predict the city hero's final score from the previous level scores. Finally, the last part of the chapter discusses the classification model.

3.1 Datasets

The dataset was provided by the study done by Antoine Coutrot (Coutrot et al., 2022). It was generated by the researchers themselves. They created a game called City hero quest. Where players were tested for their spatial navigation ability in a city environment. The study recorded the demographics and their performance in city and sea hero quest games. The participants in the study were mostly from the United States of America. The data for Sea hero Quest was also collected and in total 599 players' data was collected.

The dataset collected is divided into four parts as CSV files. The four parts are the Demographics of the players, Sea hero quest scores, City hero quest scores, and the entropy of the places. The Demographics part contains the "age" column which represents the age of the player, the next field is the education level which has the value "tertiary" and "secondary". The next column is "environment" which signifies the type of environment the player grew up in, it has four values in total that is rural, suburbs, city, and mixed. The mixed type represents the mix of two environments. The Street column contains the name of the street in which they grew up and the city column contains the name of the city in which they grew up. The "still live" column contains the information on whether the subject still lives in the same city in which they grew up or moved to a different city. The current city column contains the name of the city they moved to and also the current place of residence. The study also included a column that tells the type of environment the subjects moved to with the same values as the environment type. It also has a column that tells the environment of the current city but with binary values "city" and "non-city". A similar column named environment binary is used to classify the subject's city in which they grew up as "city" and

“non-city”. A gender column is also provided which signifies the gender of the player. Lastly, a subject ID column is provided which contains a unique ID given to all the players.

The second part of the dataset contains the scores of the players in the Sea hero quest game. The scores are normalized and are called zscore. The last column contains the zscore or the final score obtained by the player after completing all the levels. The dataset also has five other columns which are scores of the player in the Sea hero quest at levels 1,11,32,42 and 68 respectively.

Table 3.1 Sea Hero Quest scores dataset details

	L1	L11	L32	L42	L68	zscore
count	599.000000	599.000000	599.000000	599.000000	599.000000	599.000000
mean	163.601714	893.124107	2091.319416	1891.515601	2829.264966	-0.040168
std	10.240467	266.027965	1410.785072	919.025128	1513.196481	0.415547
min	147.576591	167.716019	5.000000	10.000000	81.213203	-0.735261
25%	158.071358	756.364220	1262.135891	1445.810330	1788.905324	-0.304801
50%	161.033154	792.334027	1679.105962	1608.456250	2622.419787	-0.124856
75%	166.033154	901.352906	2474.709254	1966.107242	3498.925182	0.084977
max	235.399174	3009.043640	13316.918985	8858.492650	10574.137547	2.639978

From table 1 it can be observed that the lowest zscore is -0.7352 and the best performance has a 2.6 score. Also, the table gives an idea of how the players have performed not only in the final level but also in the initial levels.

The third part of the dataset contains the scores of the players in the City hero quest game. The scores are normalized and are called zscore. The last column contains the zscore or the final score obtained by the player after completing all the levels. The dataset also has five other columns which are the scores of the player in the City hero quest at levels 1,2,3,4 and 5 respectively.

Like the previous table, it can be observed from table 2 that the highest score is 4.2 in the city hero quest game while the lowest score is -0.706. The whole distribution of the score can also be observed in table 2.

Table 3.2 City Hero Quest scores dataset details

	L1	L2	L3	L4	L5	zscore
count	599.000000	599.000000	599.000000	599.000000	599.000000	599.000000
mean	178.907503	823.811827	1408.538880	2011.317523	2518.588358	-0.006363
std	10.074452	238.055430	741.900431	647.020752	1285.917418	0.689873
min	174.118568	690.982371	880.501090	1414.979831	104.118801	-0.706967
25%	175.930958	724.278912	1026.129677	1647.406994	1651.161803	-0.455164
50%	176.753292	750.540917	1143.520245	1782.555411	2101.890172	-0.207621
75%	178.405779	804.706204	1488.959810	2107.618325	3009.795591	0.200291
max	319.618046	3719.044078	9112.265440	6124.776388	12328.193120	4.252410

Table 3.3 Entropy dataset details

	entropy_adjusted
count	595.000000
mean	2.875997
std	0.532062
min	0.693147
25%	2.529039
50%	3.026792
75%	3.313682
max	3.558401

The last and the final part of the dataset contains the entropy values of a location. Entropy is the value given to a location based on several factors. Factors such as the unique streets in location, number of turns, deviations from 90, and many more factors were considered(Coutrot et al., 2022). It also has the address of the players. With the address and considering the factors, every city is given a unique entropy value. The higher the value of entropy the more complex the city, similarly the lower the entropy the less complex the city is. Lastly, it also has a column that contains the environment type of the address, it takes values such as rural, city, suburb, and mixed, similar to the first part of the dataset. From table 3 the highest value of the entropy is 3.55 which is the most complex city while the lowest value is 0.69 which is the least complex city to navigate.

3.2 Experimental Design

In this section the steps taken to achieve the goals of the project are discussed.3.2.1 is about how the exploration data analysis was conducted. From 3.2.2 to 3.3.5 regression models and their implementation is discussed. 3.3.6 is about classification models.

3.2.1 Exploratory Data Analysis

To begin exploratory data analysis and build models, the essential libraries had to be imported, and data preprocessing needed to be done. This process is also known as extract, transform and load. The first step was to import all the basic and essential libraries such as sklearn for building models, matplotlib for plotting graphs, and pandas for getting and manipulating the dataset. The dataset had no missing values so there was no need to handle it by using missing value techniques. Categorical encoding and normalization needed to be done for building models which will be discussed later as they were not necessary for exploratory data analysis.

The previous studies(Coutrot et al., 2022) concluded that men perform better than women and entropy of the city is correlated with the performance (Coutrot et al., 2022) and education level and performance are correlated(Coutrot et al., 2022) and finally, the players with lower age perform well, while older age people perform relatively poor compared to the younger population. To visualize the performance of males relative to the entropy of the city they were born in, the indices of the male gender in the demographics was taken and used to find the values of entropy and zscore at those index values in the entropy dataset and sea hero scores dataset. To compare the female performance relative to the entropy, similar steps to the male scenario was taken but the female index was searched from the demographics dataset. In the study (Coutrot et al., 2022) they concluded that education level is correlated to the performance in the game, they found that the higher the education level obtained, the better the performance. So to replicate this male and female index was obtained from the demographics dataset with education level secondary, and the same thing was done for tertiary. A graph was plotted to see the distribution of the performance for both education levels. Also, the study (Coutrot et al., 2022) found that age is inversely correlated to performance, the younger populations perform better compared to the older population and the performance decreases over time. To test this finding the first step was to classify the age groups and get the index of players for each age group from the demographics dataset. Then get the zscores for each age group using the indices found in the step before, and

Finally, plot a bar graph for the average performance of the age groups. These were the steps taken to replicate previous studies conclusions.

The first insight explored in this project which had not been explored before was to check the difference in scores for players who have moved to a different city to the players who have remained in the same city. This was achieved by taking the indices of players and making two data frames for both values that are players who have and haven't changed the city. After getting these indices, the corresponding z scores were obtained. Then scatter plot was used to plot the values.

Following that, the scores of City hero quest and Sea hero quest were plotted on a graph to compare the performance of players in both games. To plot this graph the zscores of both the games were taken. To further explain, the indices of only males and females was taken from the demographics dataset, and the corresponding Z-scores was collected from both the sea hero score and city hero score dataset, Which was further used to plot and compare the performance of male and female for both scores. To observe the performance of the sea hero quest and city hero quest among different age groups, the indices of players were collected for each age group and the corresponding zscores were obtained and plotted for each age group.

The next step in the project was to analyze the city hero quest performance. The performance of the males and females is compared using a scatter plot. This was done by again taking the indices of male and female players and getting the corresponding zscore and plotting them. To analyze the education level and performance of the player, a similar approach to that of the sea hero quest analysis was used but instead of sea hero scores, city hero quest scores were obtained. Finally to analyze the age group and their performance in the city hero game the age groups was classified and the index of players for each age group from the demographics dataset was gathered. The zscores for each age group was obtained using the indices found in the step before, Finally, it was visually represented in a bar graph.

3.3.2 Multivariate Regression model

One of the main goals of the project is to build a regression model that predicts the final score from the previous level's scores. Here the following regression models was implemented Lasso, multivariable linear regression, and ridge regression. The first step was to preprocess the data for regression. The dataset CHQ scores had to be split into two parts,

those are target variables or the zscore and the independent variables which are the scores from levels 1,2,3,4, and 5. Once that was done the data was split into training and testing sets. The split used was a 70-30 split which is 70% training and 30% testing of data.

Features were selected by using statistical analysis. For example, p-value, r squared, prob t statistic, and so on were used to select the independent values and build the regression model. The figure below shows the graphical representation of the relationship between the dependent and independent variables.

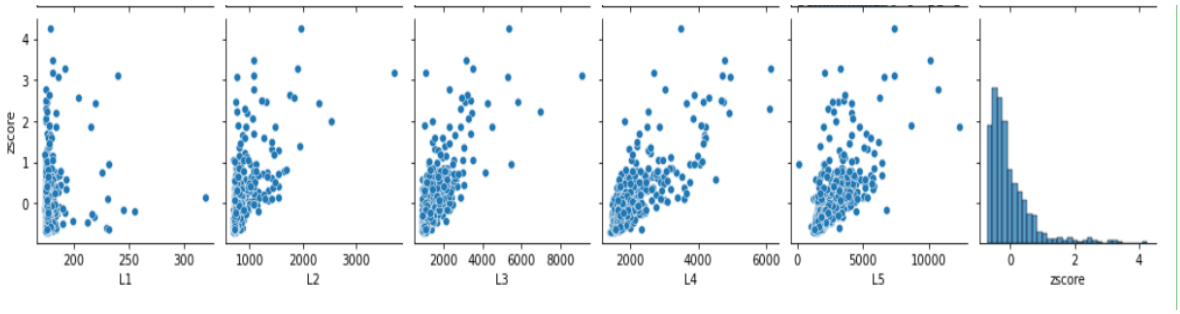

Figure 3.1 Relationship between Levels 1,2,3,4,5 and the zscore

First, the statistical values were retrieved for the variables from statsmodel library. It can be seen in the code snippet below. Using the below block of code the statistical values related to the independent variables were retrieved.

```
## Getting statistical insights for the variables

chq_stat=CHQ_zscore.copy()
chq_feat=CHQ_zscore[["L1","L2","L3","L4","L5"]]

chq_tar=CHQ_zscore[["zscore"]]

mod=smf.ols(formula='zscore~  L1+L2 + L3 + L4 + L5', data=chq_stat)

res1 = mod.fit()

print(res1.summary())
```

Figure 3.2 Code snippet for getting statistical insights

The first thing that was observed from figure 3 was that all the variables have a 0 p-value which is less than the usual significance value of 0.05(Frost, 2022). It signifies that there is a non-zero correlation between the independent and dependent variables(Frost, 2022). The coefficient value represents the change of the dependent variable given a one-unit change in an independent variable. The higher the coefficient does not mean that the independent variable carries more importance(Frost, 2022). In this case, the independent scores are not normalized scores, while the dependent scores are normalized. That is, a large change in

the independent variable will result in a small change in the zscore. Hence the coefficient is not considered for deciding which features to consider.

	coef	std err	t	P>|t|
Intercept	0.2606	0.038	6.855	0.000
L1	-0.0164	0.000	-74.353	0.000
L2	0.0011	9.24e-06	117.300	0.000
L3	0.0003	3.53e-06	92.519	0.000
L4	0.0004	4.07e-06	99.310	0.000
L5	0.0002	1.79e-06	112.222	0.000

Figure 3.3 Statistical information for independent variables

R-squared:	0.995
Adj. R-squared:	0.995
F-statistic:	2.226e+04
Prob (F-statistic):	0.00

Figure 3.4 Statistical information for independent variables continued

The next two measures analyzed for building the models were R squared and F-statistic. R2 tells how much percentage of the dependent variable can be explained by the independent variable. Here, 99.5 % variation in Y can be explained by X. The next metric used to build the model is the prod F statistic which depicts the probability of a null hypothesis is true. Since the value is 0 for this metric it implies that the overall regression is meaningful.

The final step was to check the value of R-squared with different variables, as the associated change in R-squared when new variables are added represents the model performance solely based on the newly added variable as the previous variables have been accounted for. That is It represents the portion of the goodness-of-fit that is specific to each independent variable(Frost, 2022). The figure below shows the value of R-squared values for considering different levels. The model with levels 2 to 5 as independent variables has a very similar R-squared value compared to considering all the levels as independent variables for the model. Figure 5 shows the R2 value of the model when different levels are considered as independent variables. It is apparent from the graph that the R2 value for models with levels 2 to 5 as independent variables performs well along the levels 1 to 5 as independent variables. In the plot, ‘L1” represents level 1 as the independent variable, and “L1-L5” represent all the levels from level 1 to 5 as independent variables.

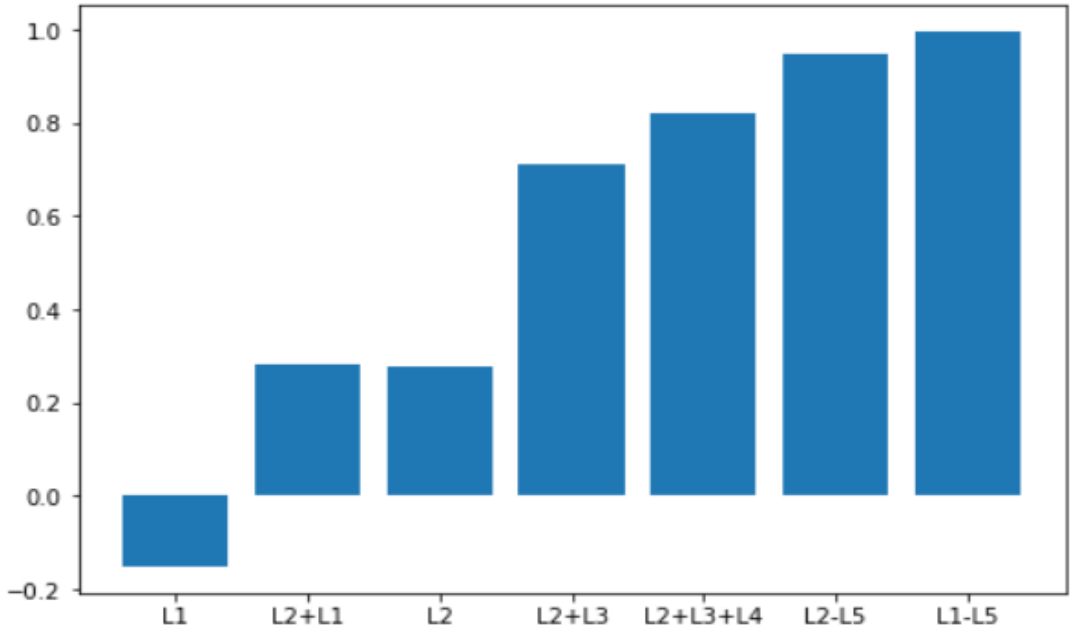

Figure 3.5 R-Squared values when different independent variables are considered (X-axis= Levels considered, Y-axis= R2 value)

Then using the library sklearn, the linear regression model was imported and an instance was created. The training data was fed to the model. After training, the model was fed with independent variables testset. The model's predicted target values and these values were compared to the original target test set values. From these comparisons, mean absolute error, mean squared error, and root mean error were calculated.

3.3.3 Lasso Regression Model

Lasso is penalized linear regression. Lasso Regression tends to make coefficients to zero if the particular variable is not important for the model by using a penalty. A hyperparameter is used called "lambda" that controls the weighting of the penalty to the loss function. lambda is called the "*alpha*" argument when defining the class.

The default value is 1.0 or a full penalty, while 0 is no penalty which is the same as regular regression. From Figure 6, it can be observed that when alpha is increased even the mean absolute error is increased. Lasso regression aims to increase bias and decrease variance. By decreasing mean absolute error when increasing penalty term the model is moving away from the predictor that has the lowest bias.

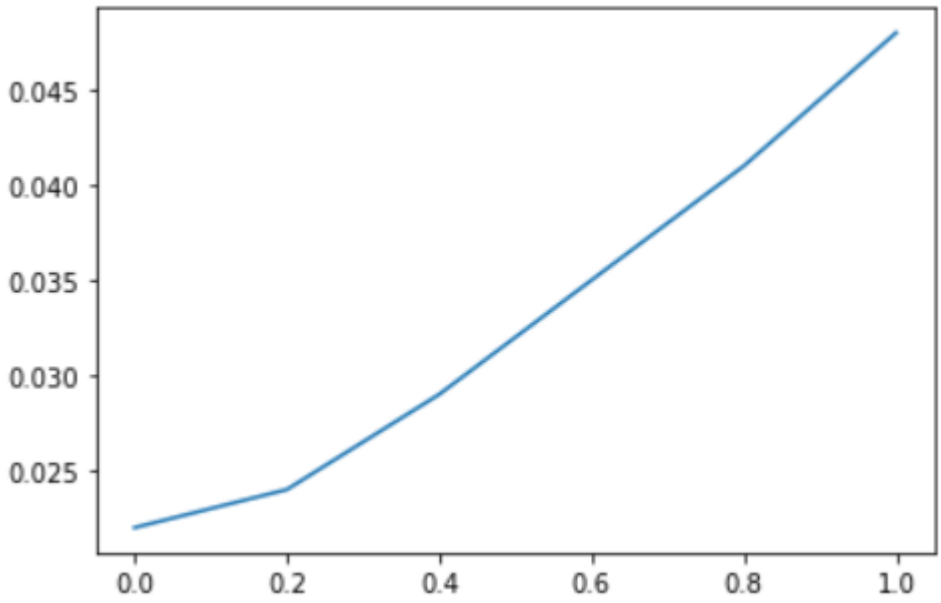

Figure 3.6 Mean absolute error vs Alpha value for Lasso regression
(X-axis=alpha, Y-axis=MAE)

Then after the analysis, using the library sklearn, Lasso regression model was imported and an instance was created. The training data was fed to the model. After training, the model was fed with independent variables testset. The model's predicted target values were compared to the original target test set values. From these comparisons mean absolute error, mean squared error, and root mean error was calculated.

3.3.4 Ridge Regression Model

Similar to Lasso, Ridge is a penalized linear regression. And a hyperparameter is used called "lambda" that controls the weighting of the penalty to the loss function. A default value of 1.0 will fully weigh the penalty; a value of 0 excludes the penalty. The lambda is called the "alpha" argument when defining the class.

The default value is 1.0 or a full penalty, while 0 is no penalty which is the same as regular regression. From figure 7, the mean absolute error is observed for different alphas and it can be observed that the mean absolute error remains constant except for 0.8. But for experimentation purposes, the alpha is taken as 1 for the model.

The model Ridge regression was called using the sklearn library. Which was fed with the same training samples as the linear regression model. The next step was to predict the values using the independent variable test set and compare these values to the actual target

values test sample. The comparison provides values for Mean absolute error, mean squared error, and root mean squared error.

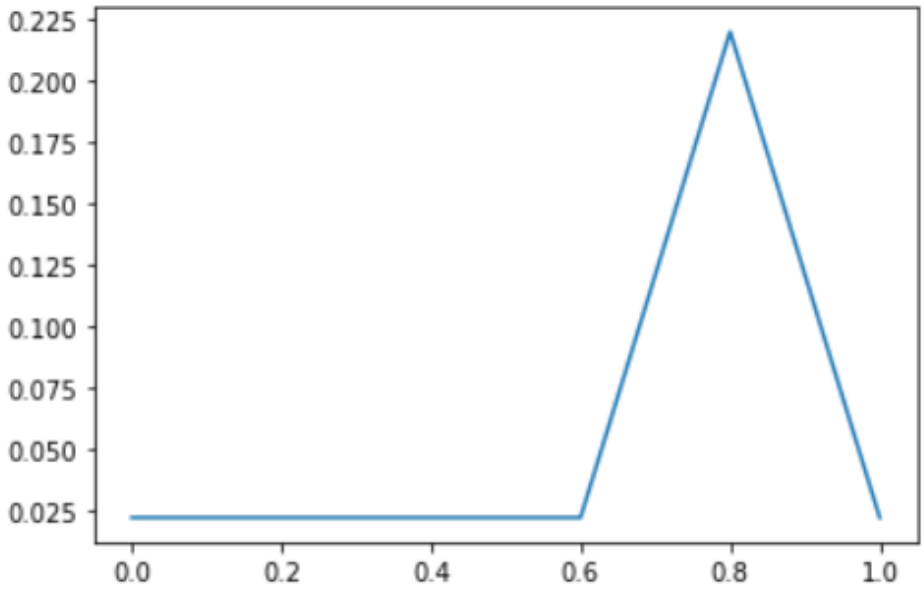

Figure 3.7 Mean absolute error vs Alpha value for Ridge regression (X-axis=alpha, Y-axis=MAE)

3.3.5 Regression Overview

The regression model was picked to build the prediction model as the target value is continuous and so were the independent variables. Also levels 2,3,4 and 5 were chosen from the model as independent variables. There was some correlation between the levels and the final score this is also a partial reason behind choosing regression models. The performance of different models is compared in the results section.

Different regression models were built and their performance was compared and the best performing model for prediction was picked. The first part was to build a multivariable linear regression. Then build regularized models such as lasso and ridge to compare the performance difference between regularized and multivariate regression. Multivariate regression performed well and to avoid the problem of overfitting, Lasso and ridge were implemented. That is Lasso and ridge regression were built to avoid any possible overfitting of the multivariable regression model. As discussed in the results chapter the performance between regularized and multivariate is similar and that implies that the multivariate model may not be overfitting and the features with the most impact on the target variable are picked.

Lasso and Ridge each have their own advantages, the common one being that they both avoid overfitting the model. Ridge regression helps reduce only the overfitting in the model while keeping all the features present in the model. It reduces the complexity of the model by shrinking the coefficients whereas Lasso regression helps in reducing the problem of overfitting in the model as well as use automatic feature selection.

3.3.6 Classification models

There were other experiments done too, that is classifying the scores into good performance and bad performance. Correlation data and exploratory data analysis were the basis for picking features for the model. Based on the correlation between the performance and the features such as age, education, gender, and environment, features were picked for the classification model. It was also concluded in this study(Coutrot et al., 2018) that the performance in the game is directly correlated to education, gender, and environment and inversely correlated to age. For this model the preprocessing included encoding. Features such as education, gender, and environment were categorical variables. This needed to be encoded to numerical values. This was done by assigning values to each category of the feature and then replacing the values in the data frame. For example, assigning male to 1 and female to 0 and then replacing the values in the column with corresponding values. The next step was to classify the scores into two categories. This was done by selecting the scores between 1 and 0 as the good category and 0 to -1 as another category. Then by importing the sklearn library Decision tree, Random forest, and Naive Bayes was implemented. The data was split into testing and training just like in the case of regression. The models were made to predict the values from the independent variable test set and compared to the actual target test set values. Using these comparisons accuracy, precision and recall were calculated.

Naive Bayes was picked, as the outcomes are two and naive Bayes is suited for dual outcomes. Naive Bayes calculates the probability of an outcome and this is helpful in classifying the performance of a player. Rather than finding whether the player's performance was good or fair. They can find out the likelihood of which category they fall into. The decision tree and the random forest was picked as they are very good classifiers and can a handle variety of classification problems. Another advantage of the decision tree is, from the studies (Coutrot et al., 2018)it is already known that age, education, and gender play a role in performance so these could be easily used to build a tree model and help in classifying.

3.3.7 Conclusion

In this chapter, the first section discusses the dataset. The second part discusses how the replication of the studies was conducted and further how experimental setups were set to gain new insights into the datasets. Following this, the chapter discusses how multivariate regression was built and how the features were picked for the algorithm. The next component was how regularized regression was implemented and how the hyperparameters were chosen for the models is discussed. Finally, the chapter discusses the implementation of classification models.

Chapter 4 Results

Section 4.1 Overview

Exploratory data analysis was conducted on the sea hero and city hero datasets. Regression models were built to predict the final score of the player from the previous level score. And finally, there were some experiments conducted on classifying the performance of the players. In this chapter, the results obtained by the experiments conducted are discussed. The insights found by exploratory data analysis are discussed in section 4.2, the performance metrics of the regression model are discussed in section 4.1, and the performance metrics of the classification model are discussed in section 4.3.

Section 4.2 Results of Regression models

In this section the results of regression models are discussed. The results from Multivariate, Lasso, and Ridge are discussed and the performance is compared.

Section 4.2.1 Multivariate regression

Table 4.1 Performance metrics of Multivariate regression

	Multivariate regression
Mean Absolute Error (MAE)	0.0784
Mean Squared Error (MSE)	0.0271
Root Mean Squared Error	0.1648
R2	0.9559

A multivariate regression model was built to predict the final city hero quest zscore from previous levels. Table 4 shows the performance metrics of the multivariate model. The reasons behind using these metrics and the meaning of the metrics are discussed in the next chapter. The lower the values of error the better the model will predict. As seen from the table, the regression model is performing well in predicting the final zscore.

Section 4.2.2 Lasso regression

Lasso picks the variable with the most impact on the target variable. The Mean Absolute value is plotted against the alpha value for the model. When alpha is 0 it means there is no penalty and which is similar to normal regression while when alpha has the value of 1 it means there is a maximum penalty. From the graph below it can be seen that the model has the lowest mean absolute error for alpha values closer to 0. Which can signify that the independent variables picked in the feature engineering process had the most impact on the model and there is no need to change the features.

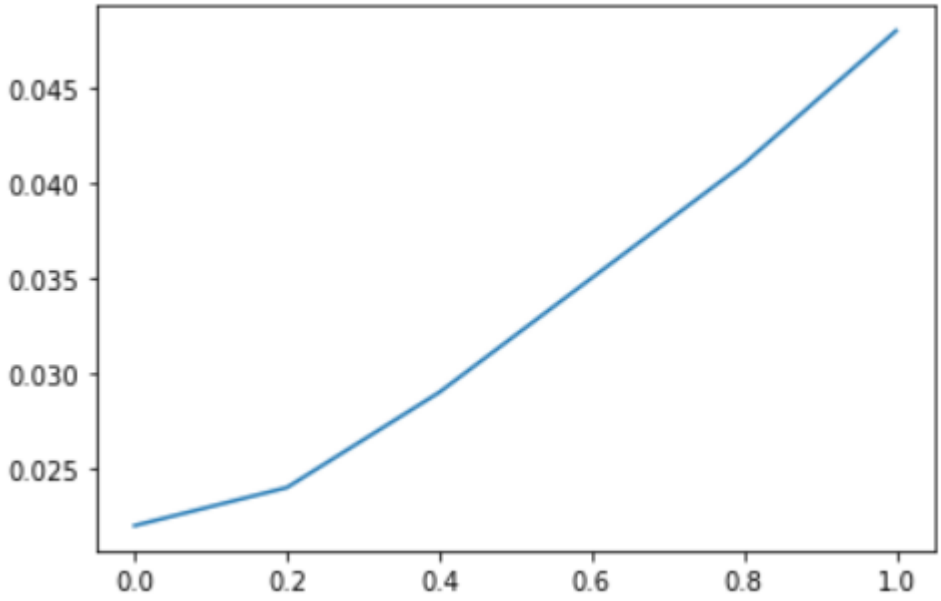

Figure 4.1 Mean absolute error vs Alpha value for Lasso regression

Table 4.2 Performance metrics of lasso regression model

	Lasso Regression
Mean Absolute Error (MAE)	0.0795
Mean Squared Error (MSE)	0.0274
Root Mean Squared Error	0.1656
R Squared	0.9554

Lasso regression predicts the final city hero quest game score from previous levels. The above table shows the errors that were calculated for the model.

Section 4.2.3 Ridge Regression

Next to compare the performance and to test if the model was overfitting, ridge regression was implemented. The Mean absolute error remained relatively constant for all the alpha values. Which can be seen in the plot below.

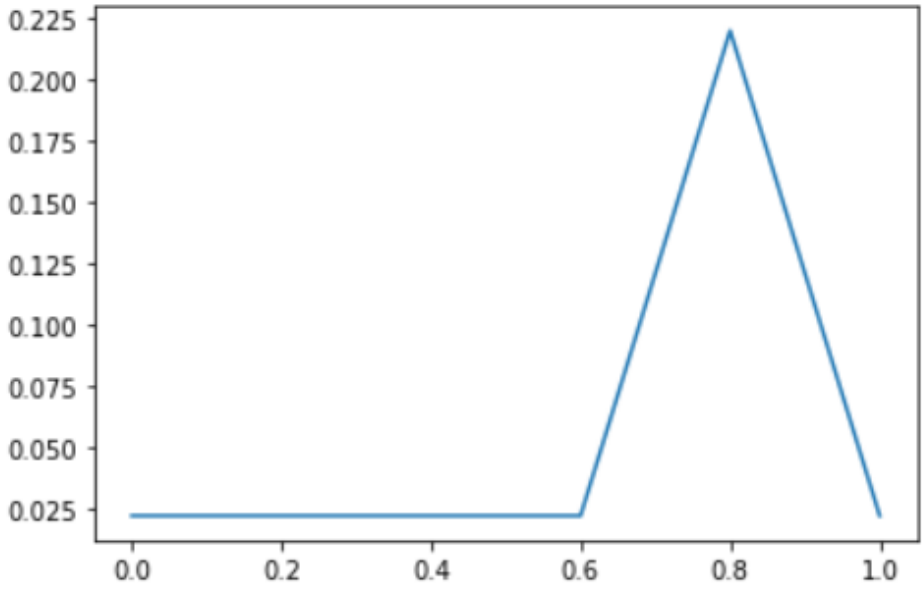

Figure 4.2 Mean absolute error vs Alpha value for Ridge regression

This can signify that the model is not overfitting for multivariate regression since the mean absolute error remains approximately the same for ridge and multivariate regression even with different values of alpha.

Table 4.3 Performance metrics of Ridge Regression

	Ridge Regression
Mean Absolute Error (MAE)	0.0784
Mean Squared Error (MSE)	0.0271
Root Mean Squared Error	0.1648
R- Squared	0.9559

Similar to the other regression models, the ridge regression model predicts the final CHQ score from the previous levels. Table 4.3 shows the performance of the ridge model with alpha 0.2.

Section 4.2.4 Comparison

Table 4.4 Comparison of Multivariate, Lasso, and Ridge models

	Lasso	Ridge	Multivariate
Mean Absolute Error (MAE)	0.0795	0.0784	0.0784
Mean Squared Error (MSE)	0.0274	0.02715	0.0271
Root Mean Squared Error	0.1656	0.1648	0.1648
R-Squared	0.9554	0.9559	0.9559

It can be observed that all the models perform very similar. With a very small difference between the models. It can be observed that there is an insignificant difference between the mean absolute error from lasso to multivariate.

By implementing the ridge regression and finding out very similar error values to multivariate, it can be concluded that the multivariate linear regression has a very high chance of not overfitting since Ridge is used to avoid overfitting and the performance is very similar to multivariate regression. Lasso eliminates certain variables to improve the performance, but in this experiment, the multivariate has a very similar performance to lasso which signifies that the variables picked for multivariate have a high impact on the model.

The error values for all the regressions model are very similar that is the difference in values are quite insignificant.

Section 4.3 Results of exploratory data analysis

The first analysis done on the dataset was on the correlation between entropy and city hero score.

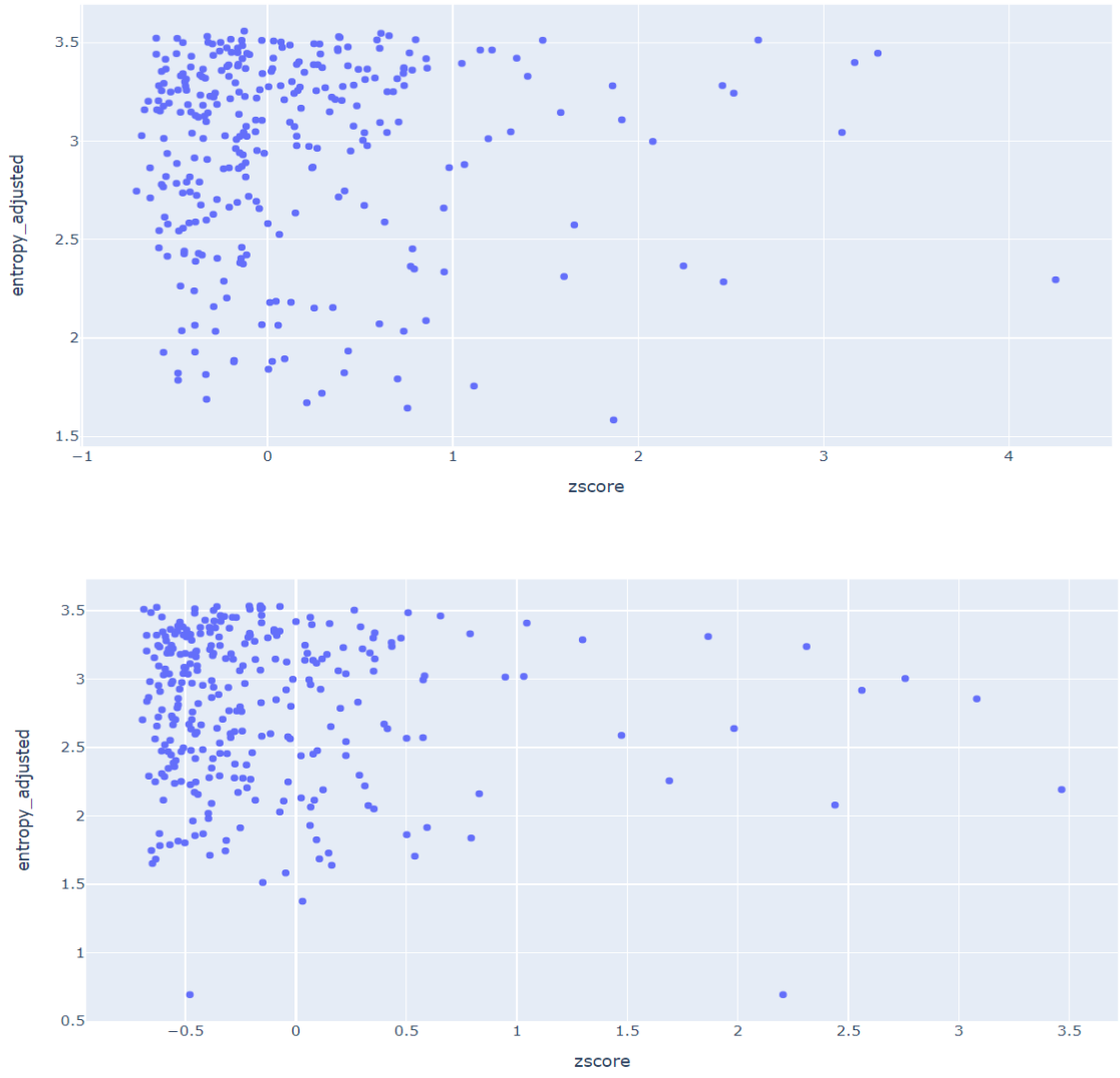

Figure 4.3 (top) Female- Entropy vs City hero score (below) Male- Entropy vs city hero score

The correlation between male player's entropy and their performance is -0.75, which means that the higher the entropy the lower the scores. It can be observed from figure 8 that when entropy is considered there is more variance in the performance of females when compared to males. The research was done for the sea hero quest but city hero quest scores and entropy were not compared before.

The second analysis was done to get insights into the education level and the performance of the city hero score.

	zscore
count	89.000000
mean	0.087411
std	0.766394
min	-0.678939
25%	-0.388465
50%	-0.142678
75%	0.401114
max	4.252410

	zscore
count	211.000000
mean	0.142371
std	0.718154
min	-0.706967
25%	-0.353357
50%	-0.064839
75%	0.398240
max	3.293040

Figure 4.4 (left) Female City hero quest zscore with secondary education. (right)Female city hero quest zscore with tertiary education.

It can be observed from figure 4.4 that the female players with secondary education score lower than players with tertiary education. The reason might be because of age but there is no data point to back this. It can also be noted that the overall performance of players with secondary education is higher compared to players with tertiary education. Similar trends can be observed even among male players. Male players with secondary education perform better than male players with tertiary education. This was proved for the sea hero quest, but the same holds true for the city hero quest by this project's analysis.

	zscore
count	170.000000
mean	-0.069230
std	0.735013
min	-0.687863
25%	-0.485843
50%	-0.335396
75%	0.065339
max	3.465768

	zscore
count	129.000000
mean	-0.231491
std	0.401631
min	-0.694041
25%	-0.533377
50%	-0.318836
75%	-0.041631
max	1.983105

Figure 4.5 (left) Male City hero quest zscore with secondary education. (right)Male City hero quest zscore with tertiary education.

The next step in the analysis was to compare the performance of city hero scores and sea hero scores.

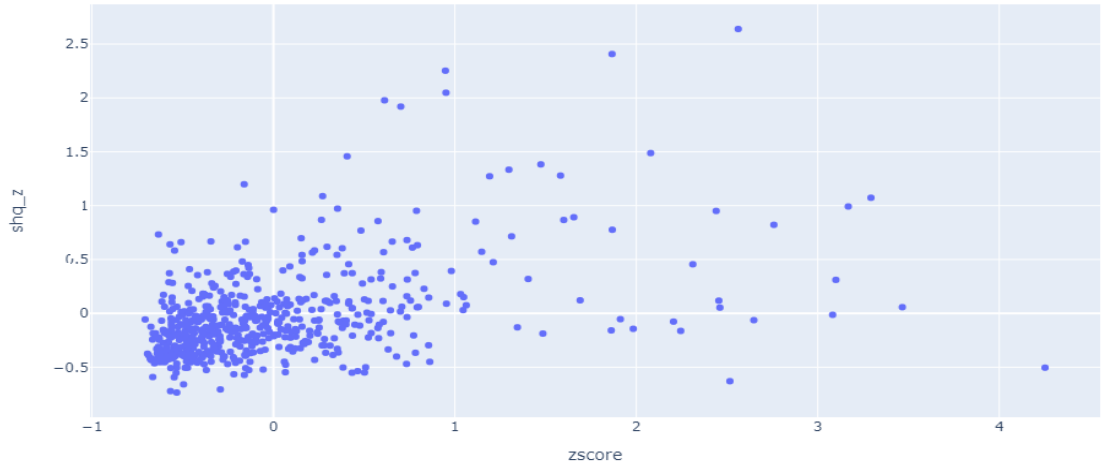

Figure 4.6 CHQ vs SHQ plot (y-axis = SHQ score, x-axis = CHQ)

It can be observed from figure 4.6 above that the performance in the city hero quest is quite similar to the performance in the sea hero quest. It can be observed that most of the player's scores are scattered inside a small area in the plot and forms a cluster. This signifies that testing the player's navigational ability from one game is quite sufficient. However, there are quite a few outliners whose SHQ score doesn't correspond to their CHQ score.

Figure 4.7 CHQ vs SHQ Female players - y-axis = SHQ score, x-axis = CHQ

It can be observed from figure 4.7 that there are quite a number of outliers among female players but mostly the performance in the city hero quest is quite similar to sea hero quest.

The correlation between CHQ and SHQ is 0.442451. It can be observed from figure 4.8 that male players have fewer outliners compared to female players. It can be said that male players are consistent with their performance between the two games. The correlation between city hero game and sea hero quest is higher for males compared to females.

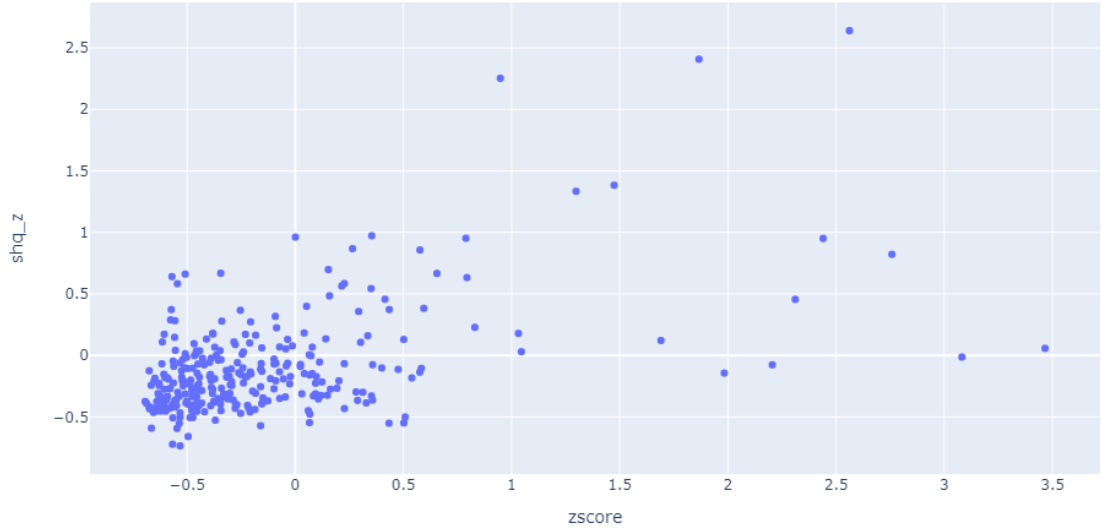

Figure 4.8 CHQ vs SHQ Male players - y-axis = SHQ score, x-axis = CHQ.

The next step in exploratory data analysis was to study the effect of age on both City hero quest and Sea hero quest scores.

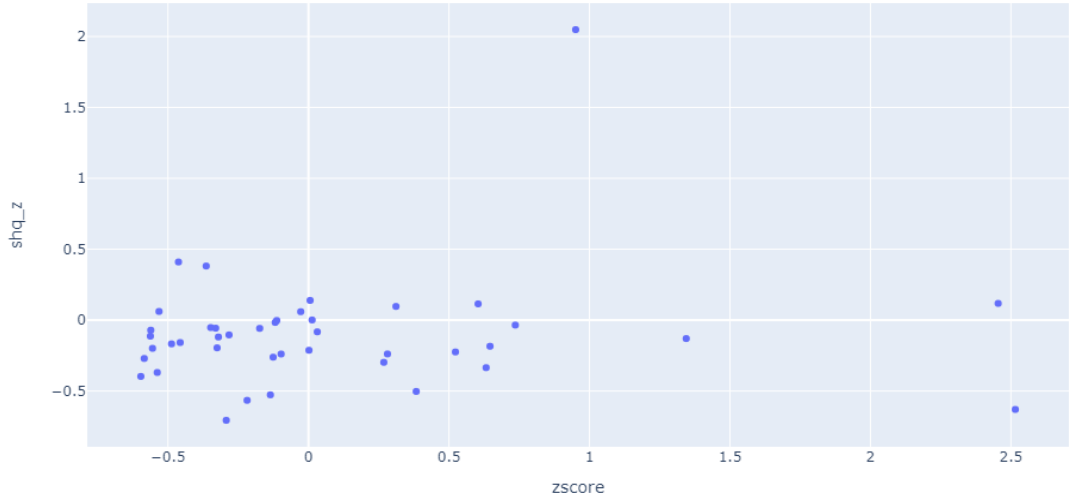

Figure 4.9 CHQ scores vs SHQ scores for players below 20 years.(x-axis=chq score,y-axis=shq score)

It can be observed from figure 4.9 that most of the players score less than 1 in city hero and less than 0.5 in sea hero quest for this age group. The correlation between CHQ and SHQ scores for the below 20 age group is 0.107115.

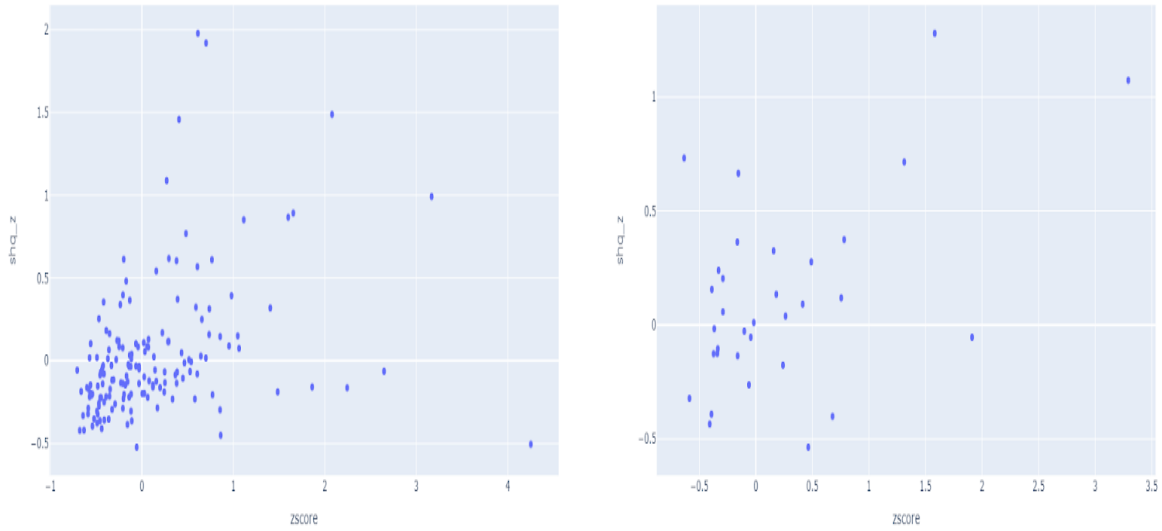

Figure 4.10 (Right)CHQ scores vs SHQ scores for players between 20 and 30 years, (left) CHQ scores vs SHQ scores for players between 30 - 40.(x-axis=chq score,y-axis=shq score)

It can be observed that for the age group 20 to 30 the range of scores is a little higher compared to the below 20 age group. For this age group, the city hero scores were below 2 and the sea hero scores were below 1. Similarly for the players in the age group 30 to 40 most of the city hero scores were below 1 and most of the sea hero quest players had scores below 0.5. But a lot of scattering can be observed in the case of the 20 to 30 age group compared to the below 20 age group. The correlation between CHQ and SHQ scores for the age group 20 to 30 is 0.35251 and the correlation between CHQ and SHQ scores for the age group 30 to 40 is 0.503348.

The sample size for the age groups 40 to 50 years and for the age group 50 to 60 years is very small. But in the limited sample size, it can be observed from figure 4.11 that most of the players score below 0.5 in city hero quest and below 0.5 in sea hero quest. Which is a significantly different range compared to the other age groups. While in the age group 50 to 60 there is no pattern as each player has a somewhat unique score, their scores show no clustering. The correlation between CHQ scores and SHQ scores for the age group 40 to 50 is 0.36371 and the correlation between CHQ scores and SHQ scores for the age group 50 to 60

is -0.282146. That signifies that for the age group 50 to 60 the higher the score in one game, the lower they will score in another game.

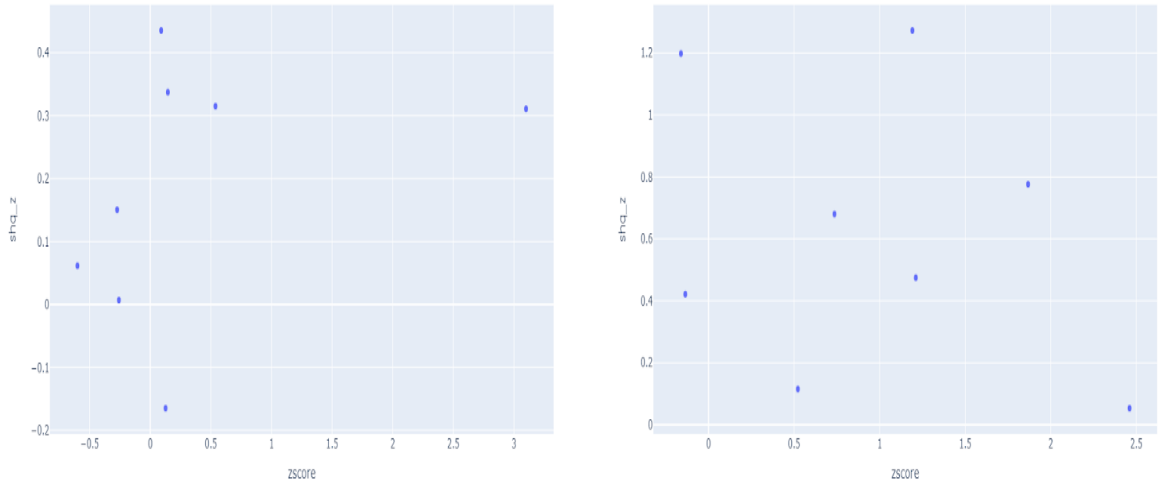

Figure 4.11 (Right)CHQ scores vs SHQ scores for players between 40-50 years, (left) CHQ scores vs SHQ scores for players between 50 - 60 years(x-axis=chq score,y-axis=shq score)

The study (Coutrot et al., 2022) collected data from players, which included the current city in which the players stay and also the city they grew up, they wanted to understand if moving to a new city would change their navigational ability. So the last part of the exploratory data analysis was to compare the performance of players who had moved and players who remained in the same place.

Figure 4.12 Still lives in the same place vs Moved to a new city(y-axis=zscore)

The scores in this plot are the sea hero quest scores. It can be observed that there is a very slight bump in the score of people who have moved to a different city compared to where they grew up. This further confirms the study (Coutrot et al., 2022) that people with experience in a variety of entropic environments tend to do well in navigational games or tasks.

Section 4.4 Results of other experiments

There were some experiments conducted to classify the scores as “good” and “fair”. In this section, the results of those experiments are discussed. This is not the main goal of the project but rather some results of the experiments reported. To get the best classification model 3 classification algorithms were implemented.

Section 4.3.1 Decision Tree

Table 4.5 Performance metrics of Decision Tree

	Decision Tree
Accuracy	0.6802
Precision	0.3913
Recall	0.4

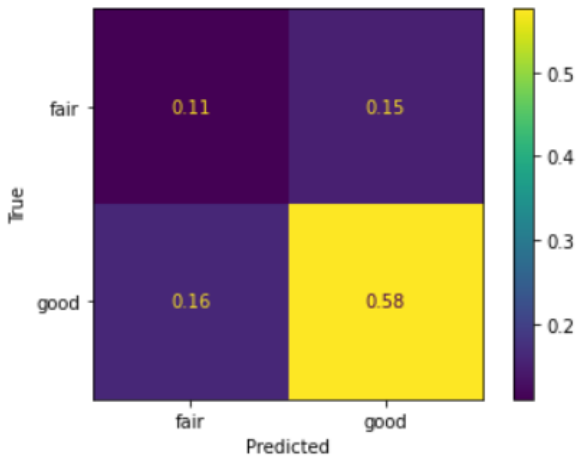

Figure 4.13 Confusion matrix for decision tree

A decision tree was built to classify the scores into “good” and “fair”. A sea hero quest performance score between 1 and 2 is considered as good while scores between 0 to -2 is considered fair. Since the scores are continuous values these categories were created for the purpose of experiments and have not been explored in the previous studies.

The model does well in predicting the “good” category but fails to produce good results for the “fair” category. The model has a higher value for false positive and false negative values. This suggests that the model is predicting it as “good” class which are actually “fair” class.

Section 4.3.2 Random Forest

Table 4.6 Performance metrics of Random forest

	Random Forest
Accuracy	0.7267
Precision	0.4687
Recall	0.333

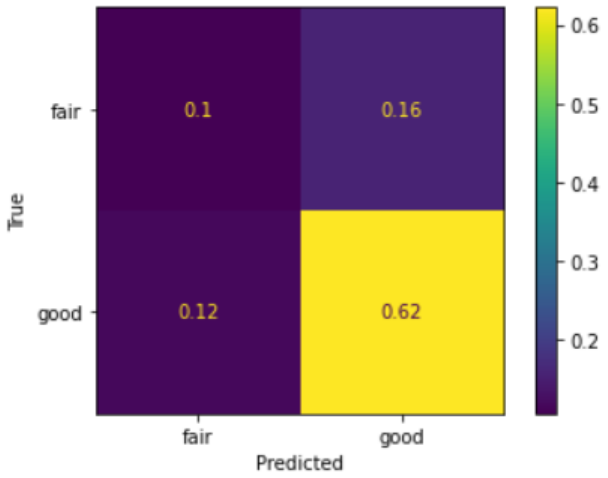

Figure 4.14 Confusion matrix for Random forest.

A random forest model was built for the same purpose mentioned above. Even though it has higher accuracy and precision it has lower recall compared to the decision tree. There could be two reasons for this; a high number of False negatives which can be an outcome of imbalanced class or untuned model hyperparameters. After changing hyperparameters and

testing the model. It is apparent that the low recall is caused by an imbalanced dataset. The model again does well in classifying “good” category but fails to classify “fair” class.

Section 4.3.3 Naive Bayes

Table 4.7 Performance metrics of Random forest

	Naive Bayes
Accuracy	0.7441
Precision	0.5294
Recall	0.2

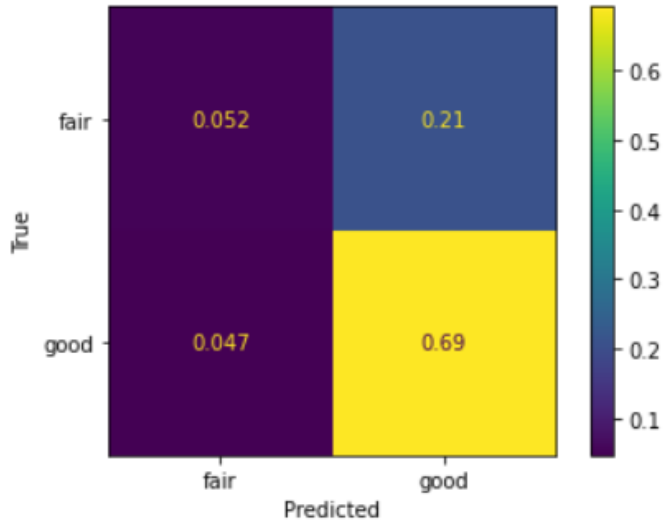

Figure 4.15 Confusion matrix for Naive Bayes.

This model has the highest accuracy and precision among all the other models but the lowest recall. Again after experimenting with hyperparameters, it can be concluded that the poor performance is due to the imbalanced dataset. This model is reasonably better than the decision tree and random forest in the prediction of the “fair” class and does well for the “good” category too.

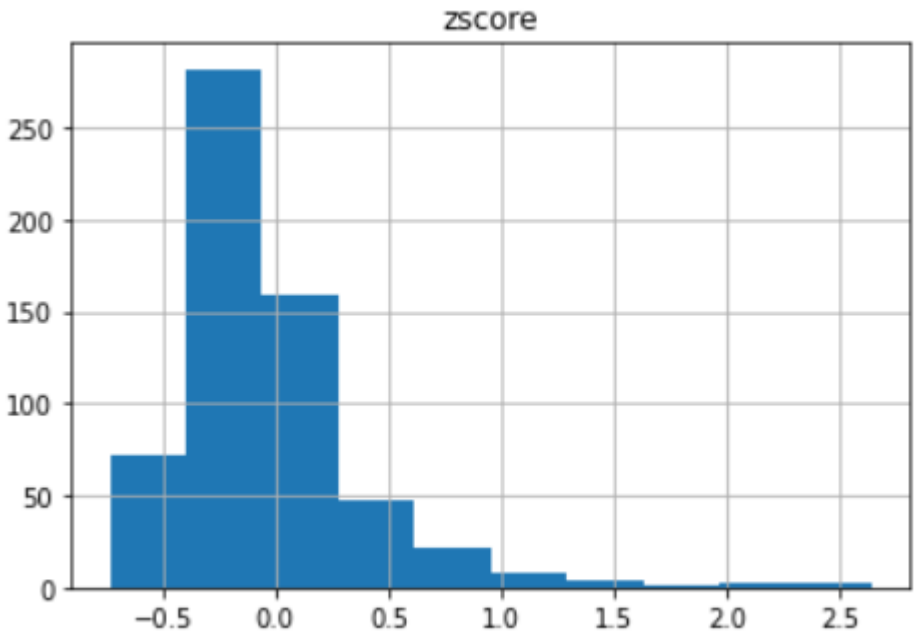

Figure 4.16 Data distribution of Sea hero quest Z-scores.(x-axis=zscore)

This is the data distribution of the zscores of the sea hero quest. It can be observed that most of the points lie between 1 and -0.5, the data is unevenly distributed for other scores. The dataset is relatively small with 599 samples and skewed towards the 1 to -0.5 range. Separating the scores into categories for classification leads to having very imbalanced data for one of the classes.

Most importantly the dataset is skewed to a certain range of scores which makes it hard to classify as there are not a lot of examples for training in the other categories. This is the reason for the low performance of the classification models. With more samples for different age groups and different scores, the model will start performing better.

Section 4.5 Results of replication

The study (Coutrot et al., 2018) concluded that age is correlated to the performance and male players perform better and players with secondary education perform better. These results were replicated by performing exploratory data analysis and the results obtained were similar.

Also, the study (Coutrot et al., 2022) concluded that entropy is correlated for sea hero quest players. This was also replicated and concluded as true. Earlier in this chapter, the relation between the city hero quest and entropy is also explored which was not done before.

Section 4.6 Conclusion

The regression model results are discussed in the first section. The metrics of linear regression, lasso, and ridge are reported with the comparison and discussion of the reasons behind the model's performance. The second section talks about the several insights from exploratory data analysis that was obtained. Insights such as the correlation between the entropy and the scores in the game, and the correlation between the scores of the city hero game and the sea quest game. The relationship between age, education, and gender of the player to the performance in the game is also established. Finally, the metrics obtained by the classification model are discussed.

Chapter 5 Validation of Results

5.1 Overview

This chapter discusses why the performance metrics were picked for quantifying the performance of models. The first section discusses why error metrics were chosen for the regression model. The second section discusses the performance metrics for classification.

5.2 Validation of regression models

Regression models approximate a mapping function from input variables to a continuous output variable. That is they predict the outputs from the inputs variables. In this project regression models predict the city hero quest zscore from the previous levels scores.

Accuracy, precision, and recall are measures of classification. Since there is no classification but rather predicting of continuous values the performance of regression is calculated as errors. Metrics such as mean squared error, absolute error, and root mean squared error are used. Errors calculate the difference between the actual values and the predicted values by the model. By measuring how close the model predictions are to the actual values. The performance of the models can be measured.

Mean squared error calculates the mean of squared differences between predicted and expected target values in a dataset. The values are squared to remove the negative values and make the error a positive error. Root mean squared error is similar to mean squared error but after calculating the mean squared error the root of the value is taken. This helps in the squaring of the values when calculating the mean squared error. MSE and RMSE punish large errors that is when the difference between the actual value and predicted values is large the subsequent square values become larger affecting the overall score. This is where Mean absolute error comes into the picture. Mean absolute error calculates the absolute error between the actual and predicted values. By using the absolute function the difference is calculated neglecting the sign of the number. All these errors calculating techniques suggest that the closer the value of the error is to 0 the better the model performs.

5.3 Validation of Classification models

Since classification models predict specific classes and not continuous values, metrics such as accuracy, precision, and recall can be used. Accuracy calculates the correct predictions to all the predictions made by the model. One of the drawbacks is it works well when there is

an equal number of samples in each class. Therefore other metrics are needed along with accuracy.

Another method is precision which calculates the total correct predictions which were actually correct, in other words, it is the ratio of true positives and all the positives. When the actual classification and the predicted classification are the same then it is known as true positives. Another method is recall which is the measure of correctly identifying true positives. There are several other methods such as ROC curve, log loss, and F1 score.

The model is not performing well because of having an imbalanced dataset. The dataset has only data from 599 players. Most importantly the dataset is skewed to a certain range of scores which makes it hard to classify as there are not a lot of examples for training in the other categories. This can be observed in the plot below.

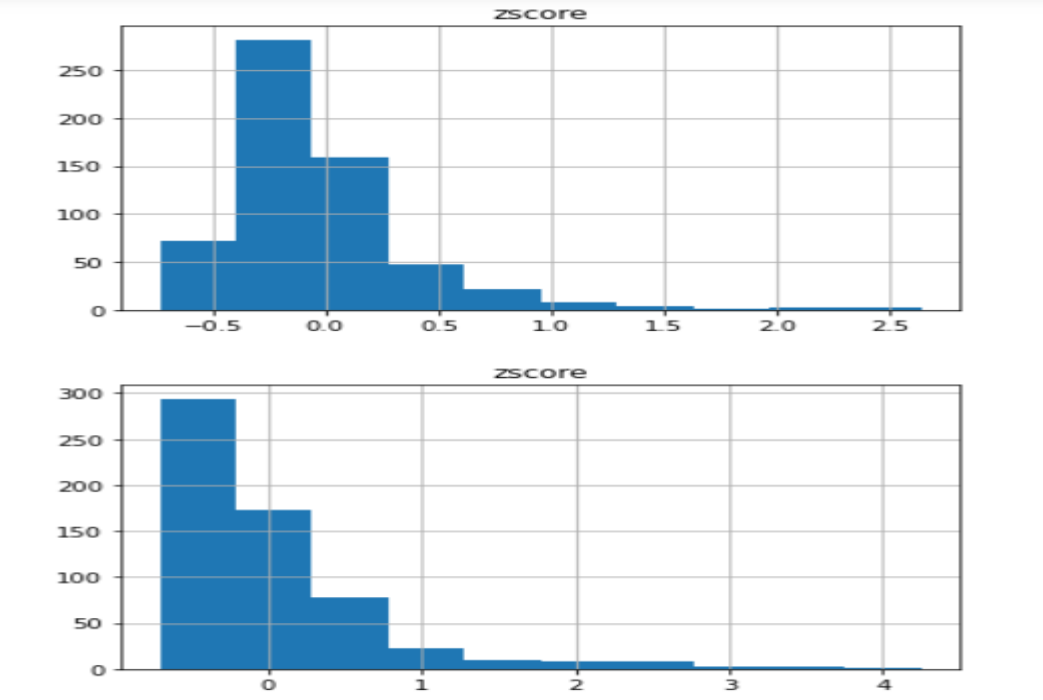

Figure 5.1 (top) Sea hero quest scores distribution (bottom) City hero quest scores

Chapter 6 Conclusion and future work

6.1 Overview

The first part of the chapter discusses the conclusion gained from this project. Conclusions such as insights gained and the performance of the model. The second part discusses future work that can be done.

6.2 Conclusions

This project focuses on getting insights from the Sea hero quest and City hero quest datasets. This project also aims to build regression models which can predict the final score of the city hero quest game from the initial levels. There are also some experiments conducted to classify the scores into two categories.

A game called Sea hero quest was created to study the navigational ability of different populations and demographics. This was done in order to build an onset of dementia detection model. The first sign of dementia is the loss of special awareness and is the reason for taking that approach by UCL. But due to the complexity of the model and the lack of data available, the model could not be built by UCL. This project helps bridge the gap from the initial idea of the onset of dementia to the actual model. This is achieved by conducting exploratory analysis and building regression models.

The insights gained from this project are:

1) There is a negative correlation between entropy and performance in City hero Quest performance. Entropy is a value assigned to a location. This signifies that the higher the entropy value of the player's location lower is the performance of the player. It is found that when entropy is considered there is more variance in the performance of females when compared to males.

2) Overall players with higher education levels perform better. This was found true for the sea hero quest but in this project, it is concluded for the city hero quest game players.

3) The correlation between the performance in the city hero quest game and the sea hero quest is 0.44. Male players have more consistent performance between the games sea hero quest and city hero quest when compared to females.

4) The correlation between the City hero quest and the Sea Hero quest game for different age groups was found.

i) Age group Below 20: Most of the players score less than 1 in city hero and less than 0.5 in sea hero quest for this age group. The correlation between CHQ and SHQ scores for the below 20 age group is 0.107115.

ii)Age group 20-30: For the age group 20 to 30 the range of scores is a little higher compared to the below 20 age group. For this age group, the city hero scores were below 2 and the sea hero scores were below 1. The correlation between CHQ and SHQ scores for the age group 20 to 30 is 0.35251.

iii)Age group 30-40: For the players in the age group 30 to 40 most of the city hero scores were below 1 and most of the sea hero quest players had scores below 0.5. The correlation between CHQ and SHQ scores for the age group 30 to 40 is 0.503348.

iv)Age group 40-50: Most of the players score below 0.5 in the city hero quest and below 0.5 in the sea hero quest. The correlation between CHQ scores and SHQ scores for the age group 40 to 50 is 0.36371.

v)Age group 50-60: In the age group 50 to 60 there is no pattern as each player has a somewhat unique score, their scores show no clustering. The correlation between CHQ scores and SHQ scores for the age group 50 to 60 is -0.282146.

5) There is a very slight increase in the score of players who have moved to a different city compared to where they grew up when they played the game.

A multivariate regression model was built to predict the final score of players in the city hero quest from the initial levels. This model had very low values for error metrics which suggests that the model is performing well. Further by implementing the ridge regression and finding out very similar error values, it can be concluded that the multivariate linear regression is not overfitting since Ridge is used to avoid overfitting and the performance is very similar to multivariate regression. Lasso eliminates certain variables to improve the

performance, but in this experiment, the multivariate has a very similar performance to lasso which signifies that the variables picked for multivariate have a high impact on the model.

Classification models were built to classify the performance of players as good and fair. Decision tree, Random Forest, and Naive Bayes were implemented. These models did not perform well. This is because of the data distribution of the zscores of the sea hero quest. It can be observed from the figure below that most of the points lie between 1 and -0.5, the data is unevenly distributed for other scores. The dataset is relatively small with 599 samples and skewed towards the range 1 and -0.5. Separating the scores into categories for classification led to having very imbalanced data for one of the classes.

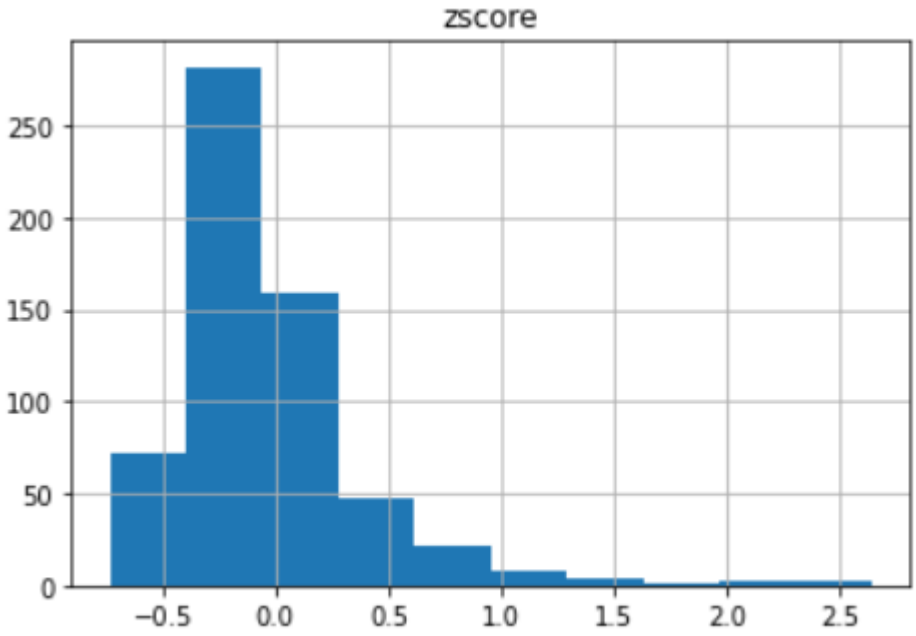

Figure 6.1 Data distribution of Sea hero quest Z-scores.(x-axis = zscore)

6.3 Future Work

This project has achieved its goals. However, the experiments conducted on the classification models did not reach the benchmarks. This was because of the imbalanced dataset available. Here are some of the suggestions for future research:

1) Gather and create a balanced dataset with players from various age groups.

2) The city hero quest data of players are from the USA. Hence gather data of players of city hero quest game from different geographic locations.

3) Build a classification model to classify the performance of players into multiple groups.

4) Build artificial neural networks to predict and classify scores.

5)Find out the relationship between different levels of the games.

6)Build a model to predict the city hero scores from sea hero scores and vice versa.

List of References

Coutrot, A., Silva, R., Manley, E., de Cothi, W., Sami, S., Bohbot, V., Wiener, J., Hölscher, C., Dalton, R., Hornberger, M. and Spiers, H., 2018. Global Determinants of Navigation Ability. *Current Biology*, 28(17), pp.2861-2866.e4.

Coutrot, A., Manley, E., Goodroe, S., Gahnstrom, C., Filomena, G., Yesiltepe, D., Dalton, R., Wiener, J., Hölscher, C., Hornberger, M. and Spiers, H., 2022. Entropy of city street networks linked to future spatial navigation ability. *Nature*, 604(7904), pp.104-110.

Seldon. 2022. *Machine Learning Regression Explained - Seldon*. [online] Available at: <https://www.seldon.io/machine-learning-regression-explained>

Li, H., Li, H., Vengateshwaran, N., Burgess, S., and Krishna, P., 2022. *Which machine learning algorithm should I use?* [online] The SAS Data Science Blog. Available at: <https://blogs.sas.com/content/subconsciousmusings/2020/12/09/machine-learning-algorithm-use/>

Frost, J., 2022. *Identifying the Most Important Independent Variables in Regression Models*. [online] Statistics By Jim. Available at: <https://statisticsbyjim.com/regression/identifying-important-independent-variables/>.

Analytics Vidhya. 2022. *Lasso and Ridge Regularization - A Rescuer From Overfitting*. [online] Available at: <https://www.analyticsvidhya.com/blog/2021/09/lasso-and-ridge-regularization-a-rescuer-from-overfitting/>.

Appendix A

External Materials

A.1 Dataset

The dataset was created by the study “Entropy of city street networks linked to future spatial navigation ability”. The dataset is contributed by Antoine Coutrot. The copyright terms permit the use of data for academic purposes. The link for the dataset is https://osf.io/7nqw6/?view_only=6af022f2a7064d4d8a7e586913a1f157.

Appendix B
Ethical Issues Addressed

The project does not include any personal or private data, and no ethical issues have been found.