MSc Dissertation

University of
St Andrews

Title: Machine Learning Based Analysis of Emotions in Medical Consultations

Major: Artificial Intelligence

Student ID: 170009479

Author: Yao Tong

Supervisor: Dr Oggie Arandelovic

Declaration

I hereby certify that this dissertation, which is approximately 12000 words in length, has been composed by me, that it is the record of work carried out by me and that it has not been submitted in any previous application for a higher degree. This project was carried out by me at the University of St. Andrews from 06/2018 to 08/2018 towards the fulfilment of the requirements of the University of St. Andrews for the degree of M.Sc. under the supervision of Dr Oggie Arandelovic.

In submitting this project report to the University of St. Andrews, I give permission for it to be made available for use in accordance with the regulations of the University Library. I also give permission for the title and abstract to be published and for copies of the report to be made and supplied at cost to any bona fide library or research worker and to be made available on the World Wide Web. I retain the copyright in this work.

Signature：Yao Tong

Date：20/8/2018

Acknowledgement

I would like to acknowledge and thank the following important people who have supported me, not only during the dissertation ， but throughout my Master degree Firstly, I would like to express my gratitude to my supervisor Dr Oggie Arandelovic, for his unwavering support, guidance and insight throughout this dissertation Secondly, I would like to thank my parents and boyfriend for their full support. You have all encouraged and believed in me. Finally, my thanks and appreciation s also go to my colleague and friends who have willingly helped me out with their abilities.

Abstract

This article details the research process of the project. The objective of the project is to detect if there are emotions are expressed in the dialogue between the patient and the doctor, and how much emotion is expressed. The result that needs to be predicted is an arbitrary value between 0 and 1. 0 means that there is no emotion in this paragraph, and 1 represents that this paragraph is full of the patient's emotions. If the predicted output is 0.5, 50% of the paragraph expressing the patient’s emotion. To achieve this goal, I used machine learning models, including linear regression, logistic regression, and support vector regression. The ultimate goal is to select and evaluate the optimal model. Through experimentation and analysis, the logistic regression model is thought of as the best model.

Contents

List of Table

List of Figure

List of Abbreviations

The table of nomenclature is as follows:

Table 0-1 List of Abbreviation

Acronyms/ Abbreviations	Full name
HCP	Health Care Provider
VR-CoDES	Verona coding definitions of emotional sequences
ML	Machine Learning
LR	Linear Regression
LOR	Logistic Regression
SVR	Support Vector Machine
CV	Cross-validation
The window	The Sliding window
R^2	Coefficient of determination
RMSE	Root mean square error

The given dataset indicates the dataset offered by the supervisor of this thesis.

The original dataset indicates the dataset after the process of doing average.

The updated dataset indicates the dataset after the process of data balancing.

1 Introduction

This chapter firstly gives an overview of the background and motivation of the research. Secondly, the problem and the objectives of this thesis are defined in the second section. At the last, the structure and outline of the dissertation are presented in the third section of this chapter.

1.1 Research background and motivation

Much recent evidence demonstrates patients' emotions can both influence and predict health outcomes. Patients with severe depression or anxiety tend to have worse pain and health condition. Therefore, accurate access to the patient's emotional state and giving a reasonable response is critical to the patient. The expression of the patient's emotions has largely caused the health care providers' difficulty in recognizing emotions. This is because patients often express their emotions implicitly. In order to clearly define the patient's explicit or vague emotions. VR-CoDES(Verona coding definitions of emotional sequences) is proposed to treat verbally expressed emotions as “concerns” and implicit emotions as “cues”. This is a reliable system that provides a way for doctors to identify patient emotions. However, it is also very inefficient, and it also creates obstacles for emotional recognition in HCP-patient communication, which is not conducive to the development of medicine. In order to solve this problem, machine learning is applied to automatically recognize the emotions and the proportion of emotions in a paragraph.

1.2 Problem and aim statement

After summarising and extending the research background and motivation, it can be found that VR-CoDES is an effective but inefficient system, and medical professionals need to spend a lot of time identifying “cues” and “concerns" based on the system. In addition, the ability to identify clues and concerns about HCP is closely related to the understanding of the system. In other words, the identification of emotions has a lot to do with different medical professionals themselves. Furthermore, cultivating their abilities also takes much time and long-term practice and experience. Since the VR-

CoDES system appears as a standard, and the recognition of emotions is a highly repetitive and regular task, developing a machine learning algorithm to replace this process becomes a good idea.

The objectives of this project are as follows:

1) Develop a novel machine learning based algorithm for emotion detection and emotion proportion prediction in medical consultations.
2) Train machine learning models and evaluate the algorithm on a large real-world data set collected by the school of medicine, University of St Andrews.

1.3 Thesis overview

This paper includes 5 main parts. The first part is the introduction of the project. secondly, context survey shows the importance of doing this project and the related research work. Thirdly, implement methodology and evaluation methodology are given. Next, the results of training models are displayed, discussed and analysed. In the last part, the summary is provided, and future work is suggested.

2 Context Survey

In this chapter, more research work on machine learning, emotion detection and prediction have been conducted. The first section “Emotion and disease treatment” aims at emphasizing the importance of patients’ emotion recognition in treatment and patient-HCP consultations. The second section points out the core idea “machine leering” to realise the targets of the project. The last section demonstrates some state-of-the-art related research findings.

2.1 Emotion and disease treatment

2.1.1 The relationship between patients' emotion and their disease

Patients emotion can affect both health condition and help to predict health outcomes[1]. In the process of disease treatment, it often accompanies the patient's emotional changes. On the one hand, Patients mood can influence treatment effect. For example, positive patients have a relatively sanguine attitude toward suicidal treatment and produce enhanced treatment response [2]. Preservation hope is an important reason for the long-term survival of some HIV or breast cancer patients [3] [4]. Depression increases mortality for people after experiencing cerebrovascular accident [5]. In addition, emotion affects pain [6]. Optimism mood is beneficial to alleviate pain, and vice versa [7]. One of the reasons why emotions affect disease outcomes is that it influences the treatment decision of patients. For instance, the bipolar emotions of prostate cancer patients decide whether they accept this treatment [8]. Moreover, mood affects the state of the organs, such as the size of the heart, salivary secretion, gastric secretion and motility, and the count of leucocyte etc [9].

On the other hand, treatment experience leads to patient changes in mood [10]. It is common that cancer patients have mood problems [10]. In the course of treatment, their emotion appears to be unstable [11], for instance, patients complain about the new-occur symptoms [12]. Furthermore, some medicine has a side effect, including emotion disorder. For example, steroids can cause patients to be easily frustrated and angry [12].

Overall, emotions are closely linked to the disease, and paying attention to patients' mood is essential. Maintaining positive emotions is conducive to strengthening cooperation with doctors, reducing pain, controlling disease progression, and achieving good therapeutic results.

2.1.2 The necessity and difficulties of the HCPs' grasp of the patients' emotions

Based on the literature on the relations between patients' emotion and their disease, Healthcare professionals (HCPs)-patient communication is necessary for clinical medicine. It is important for HCPs to take emphasize on personalised treatment instead of reacting patients in same standardized solutions [13]. In a patient-centred communication framework, it also pointed out the importance of responding to the patients' feelings [14].

However, it is significant challenging work for HCPs to offer proper emotional service in the process of patient-HCPs communication, and emotional communication is the most demanding task for research on patient-HCPs interaction [15-18]. patients expressing emotional changes are often subtle and tend to use indirect, ambiguous and non-verbal way to show their feeling, especially for cancer patients [11]. Some patients not only show their worries and uncertainties about their medical conditions but also their employment issues, life troubles and social difficulties [11]. Therefore, accurately understanding the patient's emotions becomes more difficult. Grasping the patients' negative emotions in time helps the doctors to provide more timely and targeted responses, which is significantly beneficial to the patients coping with their disease.

2.1.3 The Verona coding definitions of emotional sequences (VR-CoDES)

Clearly identifying and recognising distress emotions is necessary for patient-HCP communication research [11]. In order to reach a consensus on the definition of cues, concerns, and the response of health providers, the VR-CoDES is launched [11]. This coding system contains two manuals, one for determining cues and concerns and another for describing the types of HCPs' reaction to patients' feeling [19]. Cue is explained as "a verbal or non-verbal hint which suggests an underlying unpleasant emotion and would need a clarification from the health provider", while concern is described as "a clear and unambiguous expression of an unpleasant current or recent emotion where the emotion is explicitly verbalized" [20]. The definitions of these two conceptions reflect the ambiguity and clarity of the patient's emotional expression and give the criteria to distinguish “cue” and “concern”. Moreover, “concern” emphasizes the present anxiety rather than the past distress, which means that specific worries being expressed in past tense should be classified as “cue”. In general, cues are more complicated and difficult to identify since cues are more likely to be imperfect and vague which need to be explored by the health provider. It is common for HCPs to have different judgments to the same turn[20].

The response of HCPs performs the functions including providing space and reducing space for further disclosure of the cue or concern. Many behaviours, such as withdrawal, looking away, reading the notes, is considered as reducing space, whereas providing space behaviour includes learning forward, nodding, warm tone of voice indicating positive emotion. From these examples, comparing with reducing space providing space behaviour tends to give positive responses and encourage patients to express their emotions. However, note that the coding system does not mark which space behaviour is more appropriate. HCPs responses are consisting of immediate response, delayed response, non-verbal behaviour, explicit response, and non-explicit response. Patients-HCP communication scene is stored in videotapes, audiotapes or text transcripts, and more non-verbal information can be detected through videotapes than audiotapes. The HCPs responses are coded involving with details. Each code indicates if the response is explicit and providing space, and the specific categories of the health provider’s behaviour. The manual lists 17 different types of codes to describe HCPs response as visible as possible. More detailed information, precise definition and encoding process are available on the European Association for Communication in Healthcare (EACH) website(www.each.eu).

This coding system is widely used in various diseases, such as dentistry [21], paediatric [22], cancer [23], and dental[24].

2.2 An overview of machine learning

2.2.1 The basic introduction of machine learning

Two reasons lead to the need for learning. First, leaning is the way to solve this problem when human expertise does not exist, or when it cannot be explained. Taking the example of identifying the spoken speech, humans can complete the process of converting the acoustic speech signal into ASCII. However, because of the different attributes of individuals, including ages, genders, and accents, different conversions for the same word occurs. In this case, we cannot explain how this task is accomplished. However, by utilizing machine learning, a large number of conversion examples are collected and learned to map these to words. In addition, not all problems can be solved by directly writing a computer program. This limitation stems from the fact that the law is learned from a great many data and experiences. One of the most effective ways to solve this kind of problems is machine learning. For example, when routing data packets through a computer network, due to real-time changes in network traffic, it is necessary to modify the paths in real time to maximize the quality of service. In this case, we need a general-purpose system that can adapt to a particular environment rather than writing a program for each situation.

The concept of machine learning was proposed by Alan Turing in 1959 and experienced a few periods of prosperity and depression. Currently, machine learning is included in artificial intelligence, usually by using statistical techniques to give computers the ability to learn from data sets rather than being explicitly programmed. The method to make machine learning has the capabilities of decisions and predictions is to train a model from a training data set. Sometimes, in order to optimize the performance of machine learning models, the example data and past experience are constantly being updated and used.

Machine learning is widely used and has made rapid progress in many areas of application. This includes areas related to Internet technology, natural language processing, image recognition, medical diagnosis, bioinformatics, stock forecasting, and more.

This project “Emotion Recognition and Prediction” is an example of combing medical diagnosis and machine learning.

2.2.2 The classifications of machine learning

Based on the different purposes, there are roughly five types of classification ways to assort machine learning: according to the data label condition, the output space, the input space (feature), and the protocol. The first two ways are introduced because the selection of models in this project referenced them, and they are the most popular classification methods.

1) According to different data label conditions

Machine learning is generally divided into two categories, supervised learning and unsupervised learning, based on the criterion whether there is feedback corresponding input in the training dataset. Supervised learning refers to learning a model from a data set that there are input data and its response in each sample. The supervised learning model is trained by fitting the output data. When the input corresponding output is only partially available, it is considered as semi-supervised learning. Another type of supervised learning is Reinforcement Learning, the feedback of the program’s actions is not lost function (for ordinary supervised learning), but the punishment or rewards. This is often used in dynamic environments such as autonomous driving and game robots.

There is still dataset when using unsupervised learning, but there are no corresponding results to the data samples. The aim of unsupervised learning is to find the structure from the inputs. The necessity of its existence is that sometimes we cannot know the classification or result of the data in advance, for example investigating the types of target customers.

2) According to different output spaces

According to the expected output of machine learning, machine learning is mainly divided into regression problems, classification problems, clustering problems. The corresponding types of models to address these problems.

The output space of the regression problem is continuous data, which differ from the classification problem (aiming at gaining a discrete value). The essence of the regression model is to predict an exact number when being given test samples.

The dataset for solving classification problem contains inputs and outputs. Inputs are classified by no less than two categories. The trained classification model assigns unknown input samples to various classes. The common application is spamming emails.

Clustering problems are also about grouping the inputs. The difference between it and the classification problem is that it trains a model with an unlabelled dataset, meaning no explicit groups. Therefore, unsupervised learning is usually used to solve the clustering problem**.**

2.3 Related work

Research in the field of emotion recognition has been popular for many years. There are many ways to express emotions, such as gestures, facial expressions, writing and speaking [25]. There are corresponding studies on these aspects [26] [27] [28]. The methods used usually combine knowledge of image processing, natural language processing [29] and machine learning [29] [30]. Emotional recognition has been studied in many aspects, such as human-computer interaction [31], computational linguistics, neuroscience, psychology and behaviour [32]. Emotional recognition is often defined as a classification problem, and the models involved include support vector machine [27], neural networks, deep learning [33]. Most emotion recognition uses supervised learning, and very few are done with unsupervised learning.

Although there are many studies in the computer field related to emotion recognition, they are all solved as classification problems or natural language processing problems.

However, the purpose of this project is to predict continuous values, therefore, it is a regression problem. Past research results cannot provide many reference and instruction. Hence, the work on emotional research is no longer introduced.

3 Methodology

In this chapter, firstly, how the data is collected is introduced. Second, the procedures of data pre-processing, including average calculation and data balance, are described in detail. Third, 3 ML models to be used in this project are proposed. Finally, on the technical side, how to implement model training, model testing, and model evaluation is illustrated.

3.1 Data collection

3.1.1 Data source

The data set was obtained by transferring the patient-HCP consultation audio recording into a semi-structured textual script used as the input of the system to be studied. There were 200 face to face and one-on-one consultation recordings, involving 91 patients and 2 therapeutic radiographers. Individual consultations took between 2 and 15 minutes. These participants were female breast cancer patients aged 28-85 (average 58 years old, standard deviation of 11.3 years) who were undertaking radiation therapy at the Edinburgh Cancer Centre in Scotland during the survey. None of the participants was known to have psychiatric conditions and could communicate normally in English. They were consenting volunteers, and the related ethical approval was granted by the School of Medicine ethics committee, University of St. Andrews.

In addition, note that the data set was gained through processing the original audio recording data, and does not contain the actual conversation content. It is directly provided by the supervisor of this thesis project. Therefore, the first author of this article does not need to provide an ethical Full ethics application to obtain the data set.

3.1.2 Dataset introduction

After deleting the duplicated files, there are 422 non-repeating files, of which 221 input data files and 221 output data files. Each input file is composed by 1000 sample sized chunks, and each chunk includes 42 features. The output data files store the corresponding outcomes, namely, if there is emotion in the chunk. The outputs are either 1 or 0, 1 indicates that there is an explicit or implicit emotion (i.e. ‘concern’ or ‘cue’), and 0 suggests that there is no emotion expression in the current chunk.

3.2 Data pre-processing

3.2.1 Basic data checking

It is necessary to check and clean the training dataset before performing operations. Common problems in data sets include repetitive data, noise, missing data and inconsistent data. In the face of the first three problems, we can delete them directly. There are many ways to deal with lost data, and it is discussed in various research articles. Inconsistent data includes the inconsistency of different attribute values, inconsistency of different unit of the same attribute, or inconsistency of different data type of the same attribute. For example, when expressing the magnitude of distance, some data use meters as the unit, some others use kilometre as the unit. When inconsistency occurs, specific pre-processing of the data is required.

In addition to handling exception data, whether non-numerical data exist need to be detected. Some data are often saved in categorical (eg. Major, nationality) or ordinal type (eg. High/low), and they are requested to modify to a number.

It can be assured that there are no abovementioned issues in this given dataset. Therefore, the related solutions are omitted.

3.2.2 Do average for the given dataset

Since the aim of the target model is to predict how much emotion accounts for the paragraph to be tested, the given data set needs to be pre-processed into a data set which is suitable for addressing the issue. The procedures are displayed by figure 3-1 and word description:

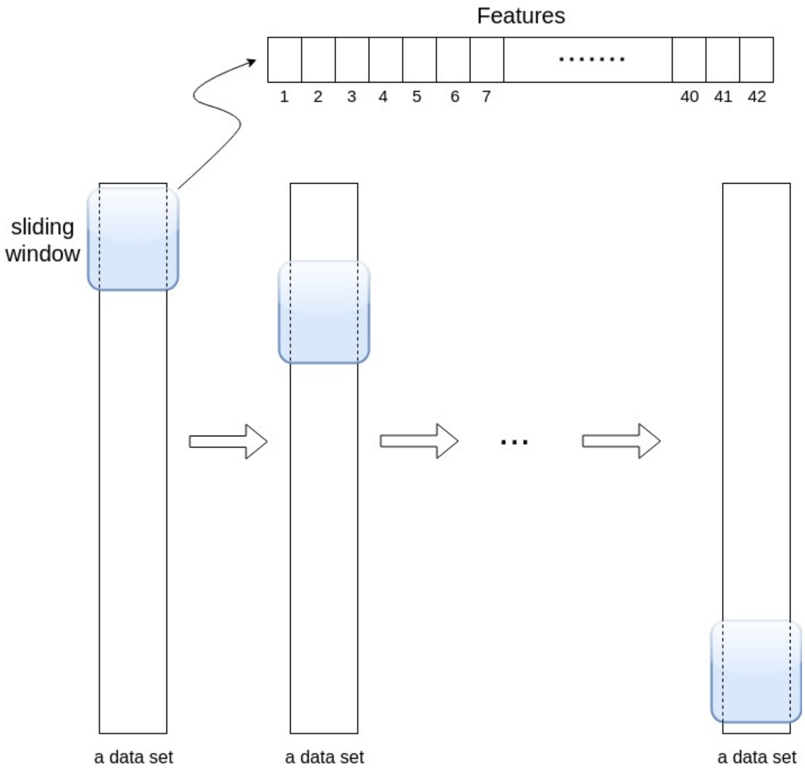

Figure 3-1 Average calculation procedure

1) Set the sliding window
    - The sliding window is initially placed at the beginning of the given data set.
    - The sliding window is used to select the data chunks in the given data set. The numbers of the framed data chunks are determined by the size of the sliding window.
    - The size of the sliding window is a positive integer which is greater than 1 and less than the numbers of chunks (i.e. 1000) of the individual data file in the given data set.
    - The window size is the first parameter of this system.

2) Average calculation
   - Calculate the average of values under each feature and the average of the outputs of the selected chunks based on the size of the sliding window, resulting in a new observation.
   - The formula of average calculation is as follows:

$$\mathrm{new\ input} = \{\mathrm{f}_1, \mathrm{f}_2, \mathrm{f}_3, \cdots, \mathrm{f}_{42}\} \tag{1}$$

$$\mathrm{f}_\mathrm{j} = \frac{\sum_\mathrm{i}^{\mathrm{i+s}} \mathrm{f}_{\mathrm{i,j}}}{\mathrm{s}} \tag{2}$$

$$\mathrm{new\ output} = \frac{\sum_\mathrm{i}^{\mathrm{i+s}} \mathrm{output}_\mathrm{i}}{\mathrm{s}} \tag{3}$$

In (1), f_1 is the average of the values under feature1, f_2 is the average of the values under feature2 and so on. In (2), f_j can be any element in the collection $\{\mathrm{f}_1, \mathrm{f}_2, \mathrm{f}_3, \cdots, \mathrm{f}_{42}\}$. Namely, j is 1, 2, 3, …, or 42, and implies the ordinal number of the column. In both (2) and (3), i represents the current position of the sliding window, and s is the size of the window. $\mathrm{f}_{\mathrm{i,j}}$ indicates the value in the position of the i row and j column in the current input data file of the given data set. $\mathrm{output}_\mathrm{i}$ is the output value in the i row in the current output data file of the given dataset.

3) Move the sliding window forward
   - After completing average calculation, the sliding window advances to select the next group of chunks.
   - How many units (or chunks) of the sliding window move forward every time is defined as the "movement way" of the window or the "separation/distance" of two adjacent windows, which is set by the operator as the second parameter of the whole system.
   - The prior constraint of the "movement way" is that it should be at most one window size and no overlap.
4) Repeat step 2), and 3)
   - For each input and output file in the given data set, repeat the above operations (the steps: 2 and 3) to get new inputs and outputs.
5) Generate the original dataset

- The new data set that completes steps 1, 2, 3, and 4 is called the original data set, and each pair of input and output in the original data set is called the "observation".
- The number of observations of the original data set depends on two parameters – the size of and the movement way of the sliding window.

3.2.3 Data balancing

3.2.3.1 The reason for balancing the original dataset

In the original dataset (the dataset that completes average calculation for the given dataset), most observations accompany with the output value in the range of 0 to 0.5, especially concentrating on 0, whatever the size and movement way of the sliding window is. In fact, it is easy to foresee this phenomenon by observing the given data set. Before doing the average calculation, 215609 chunks have no emotion, and the total number is 221000. In other words, chunks with output value 0 approximately account for 98%, which leads to most of the outputs in the original data set must be concentrated near the value of 0.

In addition, plotting output value and output distribution also evidence that there is a huge bias in the original data set. The figure3-1 displays the output value of part of observations, and the figure 3-2 demonstrates the distribution of the output value of the entire data set, which suggests that the large bias exists in the original data set.

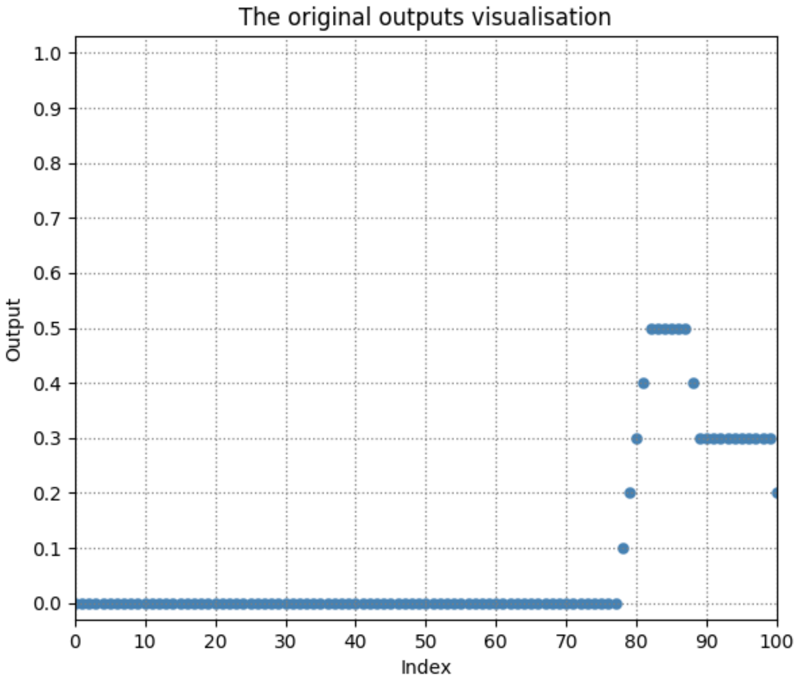

Figure 3-2 The original outputs visualisation

Figure 3-2 The graph is a scatter plot of the outputs in the original data set. The x and y-axes are named "Index" and "Output", respectively. The figure only shows the output values of the first 100 observations. Most of the points have an output value of 0, which demonstrates that there is a big bias in the data. At present, the window size is 10, and the separation of every two adjacent sliding windows is 1.

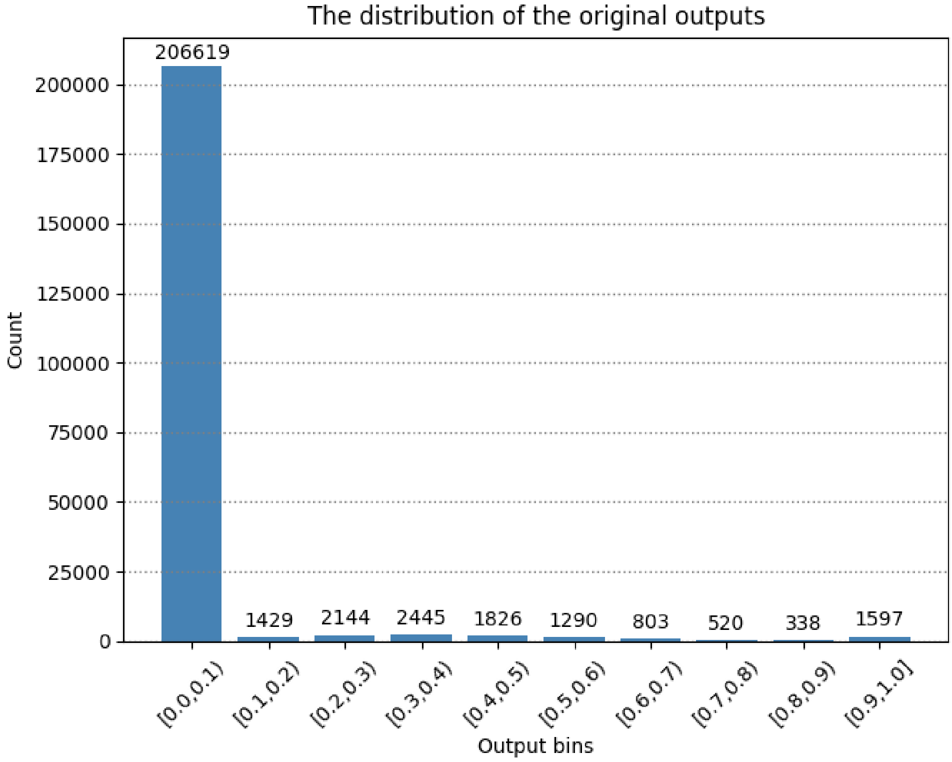

Figure 3-3 The distribution of the original outputs

Figure 3-3 The bar chart visualizes the statistical results of the distribution of the original output. It is known from the numbers of observations corresponding to different output bins that the observations with the output value 0 are the vast majority, revealing the existence of the bias in the original data set. At present, the window size is 10, and the separation of every two adjacent sliding windows is 1.

3.2.3.2 The data balance algorithms

Six data balancing algorithms were proposed to alleviate the data bias problem. They all follow the principle of removing excessive observations whose output value is 0 (representing no emotion), but the specific methods are different.

In order to reduce the bias in the original data set, five countermeasures are considered to propose five algorithms. The first aspect is to lessen the number of observations in the class where 0 is located, such as Algorithm 3 and Algorithm 5. The second method is to reduce the number of observations in the range of 0 to 0.5, such as Algorithm 2 and Algorithm 4. The third is to make the number of samples of each classification the same, such as Algorithm 1. The detailed explanation of these five algorithms is as follows;

1) Algorithm1

Ideally, the number of observations in different classifications are the same. To achieve this goal, first, compare the numbers of observations in different categories of the original dataset, and then set the minimum number to the number of different classifications of the new dataset. Finally, the corresponding number of observations are randomly taken from each category. Note that in the original data set, there may be cases where there is no observation that matches a certain classification. Therefore, the minimum value should be greater than 0 to ensure that the updated data set is not empty, and the original empty classes are not processed. The distribution of the updated data set result from using algorithms1 is shown in figure 3-4. figure 3-5, the scatter plot of these outputs, manifests the success of balancing data. The observations corresponding to these disordered outputs integrate the desired data set.

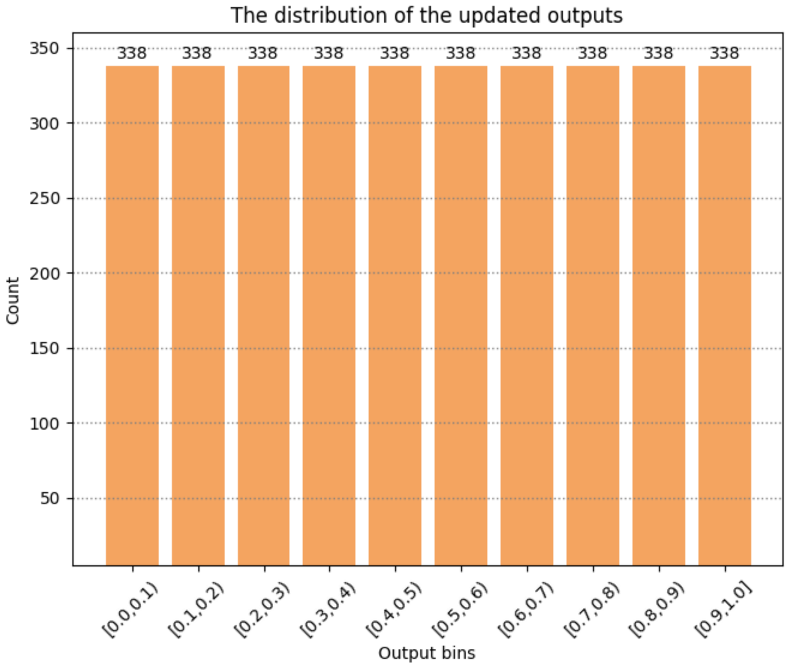

Figure 3-4 The distribution of the updated outputs

Figure3-4 In every output interval, 338 observations from the original data set are preserved. The number is obtained by running the data balance algorithm 1. At present, the window size is 10, and the separation of every two adjacent sliding windows is 1.

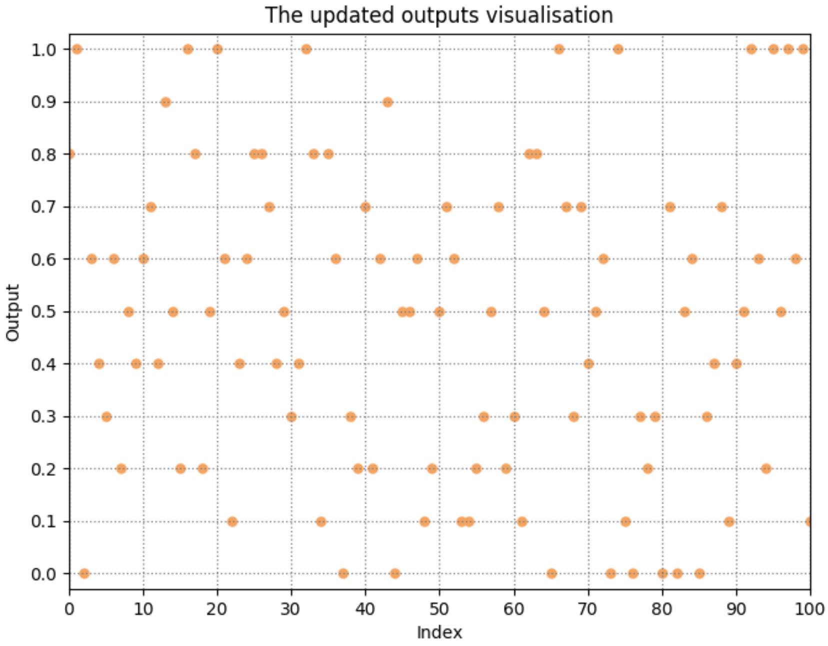

Figure 3-5 The updated outputs visualisation

Figure 3-5 The graph plots the outputs produced by completing the data balance procedure, and it only shows the output values of the first 100 observations as examples. The outputs are spread over various values in the interval of 0 to 1, and the distribution looks evenly to the naked eye. At present, the window size is 10, and the separation of every two adjacent sliding windows is 1, and data balance algorithm1 is used to generate the updated data set.

2) Algorithm2

Similar to the implementation steps of Algorithm 1, the difference is that algorithm2 is to find the median instead of the minimum. The number of observations being preserved in different classifications is the median or smaller than the minimum. This means if the number of observations in a classification is greater than the median, then only the median observations are randomly retained. Otherwise, all the observations from the classification will be preserved. The figure 3-6 illustrates the distribution of the outputs by using this algorithm.

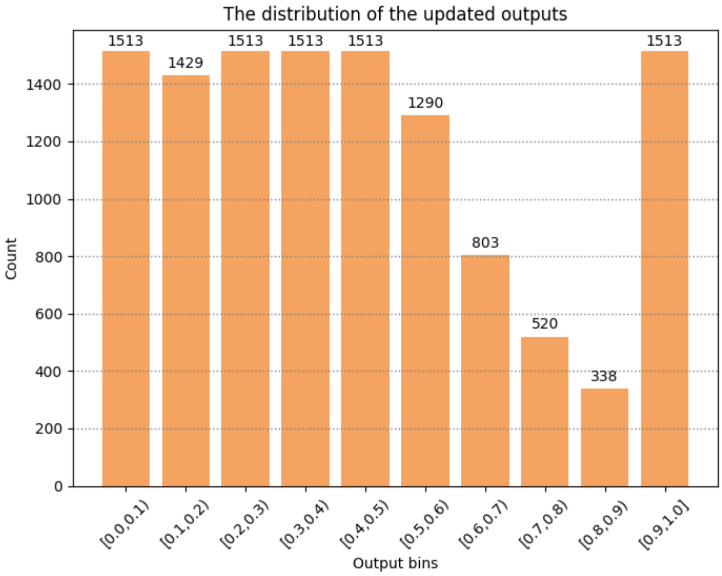

Figure 3-6 The distribution of the updated outputs

3) Algorithm3

The second maximum number is calculated through this algorithm. As it mentioned at the beginning of the section, regardless of the size or movement way of the sliding window, the class in which 0 is located has the largest number of observations. Therefore, the purpose of using algorithm 3 is to merely reduce partial observations from this class.

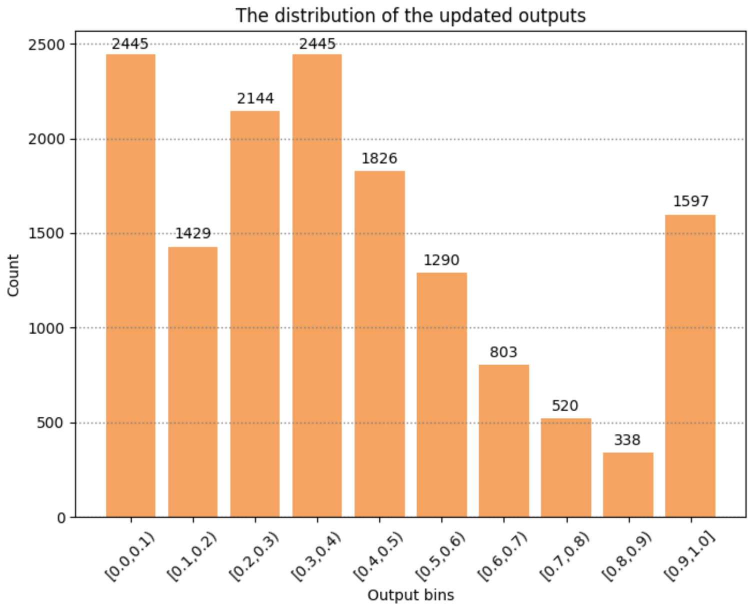

Figure 3-7 The distribution of the updated outputs

4) Algorithm4

The number of observations in a class is determined by the minimum between the two opposite classes. To explain the concept, for instance, when the size of the window is 10, and the outputs are divided into 10 categories “[0,0.1), [0.1,0.2), [0.2,0.3), [0.3,0.4), [0.4,0.5), [0.5,0.6), [0.6,0.7), [0.7,0.8), [0.8,0.9), [0.9,1.0)”, there are 5 pairs of opposite classes: [0,0.1) and [0.9,1.0), [0.1,0.2) and [0.8,0.9), [0.2,0.3) and [0.7,0.8), [0.3,0.4) and [0.6,0.7), [0.4,0.5) and [0.5,0.6). Assume that the number of observations in [0.0-0.1) and [0.1-0.2) is 998 and 345, then 345 observations in [0.0-0.1) will be retained due to the minimum is 345. This procedure is repeated until the updated data set is generated. As shown in Figure 3-8, the distribution of the updated dataset in various classes should be symmetric.

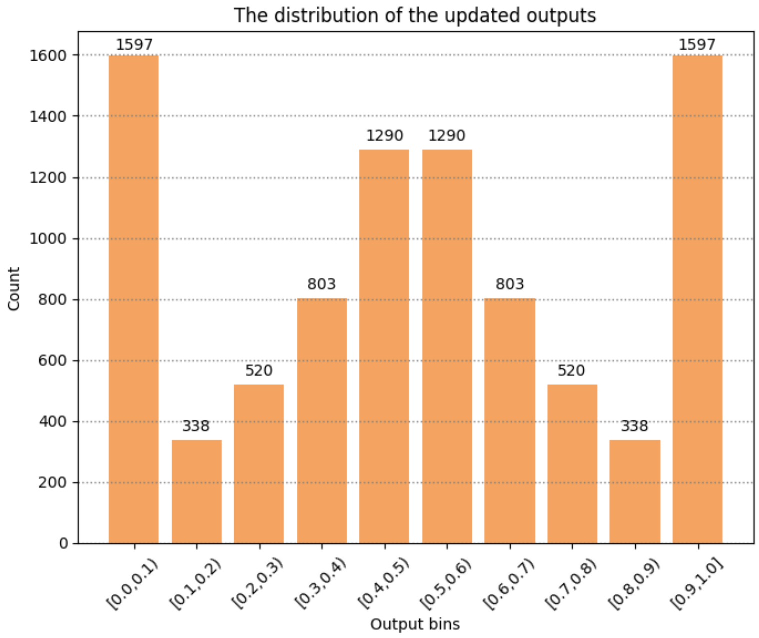

Figure 3-8 The distribution of the updated outputs

5) Algorithm5

Contrary to reducing data, another idea is to increase the data. Concretely, the number of samples per classification is set to match the number of samples of which the second largest class. For classes larger than this, the extra samples are removed. When the number of samples in a certain class is less than the value, part of the sample is randomly copied until the target size is reached.

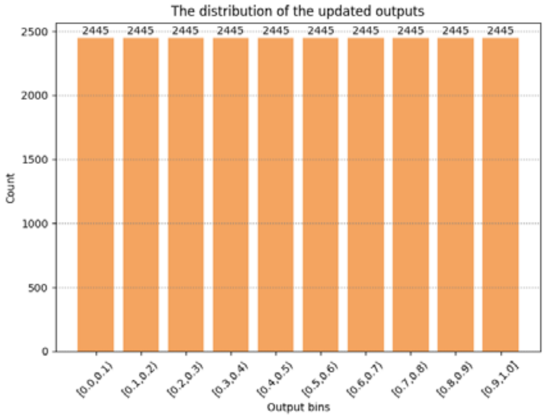

Figure 3-9 The distribution of the updated outputs

6) Algorithm6

This algorithm resembles the fifth algorithm, but the standard converts to the number of samples in each class being consistent with and the median. The median refers to the median of the numbers of samples for all categories.

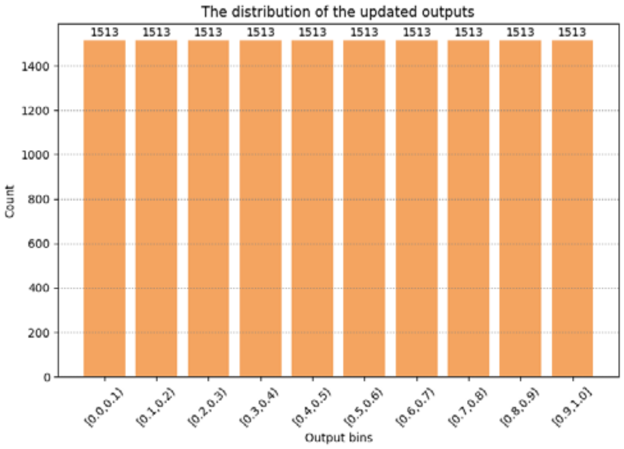

Figure 3-10 The distribution of the updated outputs

3.3 Machine learning model selection

Machine learning problems mainly include two categories: value prediction and classification prediction. According to the requirement of this project, namely, detecting if there is emotion in the text and predicting the exact proportion, it is regarded as a regression/value prediction problem. Both supervised learning models and unsupervised learning models are provided to achieve the objective. The biggest distinguish between these two types of models is whether using a labelled dataset or not. The former is based on the labelled data, and the latter learns from unlabelled data and find some structure from the data. As the data being used in this project are ladled, the related supervised learning models are determined to use. Commonly used supervised regression models involve linear regression, polynomial regression and support vector regression. It is worth mentioning that in some special cases, logistic regression can also be used to address regression problems. This experiment used this model flexibly as a novel approach. Additionally, traditional methods, linear regression and support vector regression are also applied to build models. Since LR and SVR are common models, I just introduce the novel way of using LOR here.

3.3.1 Logistic regression

The essence of logistic regression is the sigmoid function.

$$\mathrm{h(t)} = \frac{1}{1 + \mathrm{e}^{-\mathrm{t}}} \qquad (1)$$

$$\mathrm{t} = \ \beta \mathrm{X} \qquad (2)$$

$$\beta = [\beta_0, \beta_1, \beta_2, \dots, \beta_n] \qquad (3)$$

$$\mathrm{X} = \ [1, \mathrm{x}_1, \mathrm{x}_2, \dots, \mathrm{x}_n]^{\mathrm{T}} \qquad (4)$$

Here, $\mathrm{h(t)}$ is the response/output of the given t. X is the transposed matrix of the features $\mathrm{x}_1, \mathrm{x}_2, \dots, \mathrm{x}_n$. $\beta_0, \beta_1, \beta_2, \dots, \beta_n$ are the corresponding weights which are expected to be tuned. The value of these coefficients directly affects the accuracy of the model.

(1)(2)(3)(4) can be transformed into a new formula:

$$\log_e\left(\frac{\mathrm{h(t)}}{1 - \mathrm{h(t)}}\right) = \beta_0 + \beta_1 \mathrm{x}_1 + \beta_1 \mathrm{x}_2 + \cdots + \beta_n \mathrm{x}_n \qquad (5)$$

The right side of equation 5 represents linear regression, which inspires the use of logistic regression for addressing regression. Namely, the original outputs “$\mathrm{h(t)}$” are transformed into new output “$\log_e\left(\frac{\mathrm{h(t)}}{1-\mathrm{h(t)}}\right)$”.

However, it is noticed that not all logistic regressions can be used to handle the problem of continuous value prediction. The special condition of this project is that the output value ranges from 0-1, which is why this novel method can be used. In addition, in fact, we must ensure that the response value is neither 0 or 1, otherwise, the formula is undoubtedly unmeaningful. Observing the original output can be found to not satisfy this detail, so I pre-processed the source output (0 and 1) that does not meet the criteria. That is, 0 and 1 are approximately transferred to 0. 00000000001 and 0. 99999999999respectively. There are two other advantages to this setup. First, even if the size of the window reaches a maximum of 1000, it is still guaranteed that the extreme possible output 0.001 (when there is only one given output with emotion, after doing average, the output is 0.001) differs from the 0. 00000000001(which is used to institutes 0) by a high order of magnitude. Second, the number of digits after the decimal point is the largest that can make the computer calculate without errors.

3.4Implementation and Design

3.4.1 Programming Language

To implement the target of addressing the practical emotion prediction issue, Python is used to training and evaluating machine learning models. Python was designed by Guido van Rossum in the Dutch Academy of Mathematics and Computer Science in the late 1980s and early 1990s [34]. Python is a high-level interpretive, compulsive, interactive, and object-oriented scripting language[35]. The merits of Python language are easy-to-learn, high code readability, high extensibility, available to port to many platforms, having extensive standard libraries, interactive mode, interface to major commercial databases, supporting GUI, or the ability to be embedded in C /C++ [35, 36].

Python as a scripting language, omitting the compilation process of the compiled languages, such as C++, which significantly saving time and improve performance[37]. It is especially suitable for machine learning area, usually involving a great amount of calculation. Furthermore, since Python is a dynamically typed language, it does not need to declare the type of data. For machine learning that also has the potential to handle a variety of data at the same time, using Python is more time and effort saving. Moreover, Python includes many industry-renowned libraries, which offers convenience to do statistical analysis.

In this project, many common libraries, NumPy, Pandas, Math, OS, Matplotlib, Scikit-learn, Collections, and Statistics are used to implement codes.

3.4.2 Train models

3.4.2.1 Split the updated dataset

Traditionally, a dataset is divided into two parts, the training set and the test set. The training dataset is used to train a model, and the test set is used to estimate the model. The quality of the training dataset influence how good the model is. It is inaccurate and not rigorous to evaluate a model by one-time randomly dividing the data set into a 70% training set and a 30% test set. Nevertheless, k-cross validation makes up for the shortfall.

The difference between the traditional way and k-fold cross validation is that the latter groups the dataset for k times. 5-fold cross-validation and 10-fold cross validation are most commonly used to split a dataset, and many trails evidenced that 10-fold cross validation usually performs best in error estimation in many cases. Therefore, 10-fold cross validation is used in this thesis.

The description of how to split the updated dataset is as follows:

- Split data into 2 parts: the sub-dataset used in 10-fold cross validation (90%) and the test dataset.
- The first-part dataset is composed of training dataset and validation dataset.
- The test dataset is used to simulate the unknown world data which is never used to train a model. The function of it is to assess the model.
- The distinguish between validation dataset and test dataset is that validation dataset is dynamically changing in the process of 10-times splitting when

executing 10-fold validation, and every sample in the validation dataset has other 9 times opportunities to be candidates in training dataset, while the test dataset only comes in handy at the very end.

3.4.2.2 Feature selection

When collecting data, we usually collect samples and features as many as possible. Adequate features ensure avoid under fitting problems and missed fields that might have a negative effect on predicting the value. However, sometimes excessive features can also cause great computational cost and over-fitting problems. While fewer features ease interpretation. Additionally, for sake of keeping the model concise, it is necessary to ignore some features which have little influence on the output. Therefore, selecting a subset of features is commonly executed before training a model. In this thesis, optimal features are chosen in every experiment, and less relevant or irrelevant features are abandoned.

In every trial, the relative proper number and a suitable combination of features are selected. The criterion is to remove all features except the k features with the highest F-value. F-value is a popularly used feature evaluation score for regression problems. F-value is defined as the ratio of the mean regression sum of squares divided by the mean error sum of squares. The value of F is in the range of 0 to infinity.

In addition to determining the suitable scoring criteria, it is also crucial to choose the optimal number of features, that is k mentioned above. Due to some restrictions, for example, the specific meaning of the 42 features being unknown, it is significantly difficult to decide the exact value of k. An inefficient but effective solution is to traverse k from 1 to 42. Considering the total number of attributes is not very large, this solution is feasible.

3.4.2.3 Train models

The training dataset which is separated from the bigger partial sub-dataset is fitted by models containing linear regression, logistic regression and support vector regression.

The characteristics of these three models, their applicable environment, and the reason for selecting them are discussed in the previous section 3.3.

3.4.3 Test models

3.4.3.1 Model evaluation process

To evaluate a model, firstly, determine the matric. The LR, LOR and SVR models are mainly assessed by the two type of score: coefficient of determination (R^2) and root mean square error (RMSE).

(1) The interpretation of R^2 and RMSE

R^2 interprets how well the model fit the data. It gives the value that the percentage of total variation is described by the variation in x. $\mathrm{R}^2$is between 0 to 1. In this range, more approaching to 1 means the regression model is more satisfying. Although it is inferred from the formula of r that r cannot be a complex number or greater than one. But in fact, R^2 will exceed the range of 0 to 1 in some special cases of which the model fitting data is worse than a horizontal hyperplane. Simply put, if the wrong model is selected or nonsensical constraints are used by mistake, the abnormal R^2jump out. R^2 is below than zero or over than one when the equation of 1 or 2 of Kvalseth is respectively applied [38].

RMSE is another measurement score often used in statistics, expressing the average model prediction error in units of the variable of interest. The higher the value the worse the prediction results. Note that the magnitude of variance is relative to the range of and the general value of actual outputs. When different models predict different targets, the same RMSE does not mean that the two models have the same forecasting ability. For example, there are two models, model a is used to predict the height of a person (unit: feet), and model b is used to predict the age of an individual (unit: years old). Assume that the two models get the same RMSE, indicating that model does not has the as good predictive capability as model b. This is due to the range of height is relatively small, the same value of RMSE has a greater impact on height forecast.

The formulations of these two metrics are as follows:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{1}$$

$$SS_{tot} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{2}$$

$$SS_{res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{3}$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \tag{4}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{5}$$

The first 3 equations are combined to generate the mathematical definition of R^2. SS_{tot} is called the total sum of squares. SS_{res} is the residual sum of squares of residuals. These four formulas use repeated variables. To illustrate them, $\bar{y}$ is the value of the average of the total outputs, y_i is the actual output of the i^{th} the sample to be forecast, and $\hat{y}_i$ is the predicted output by the model. n is the number of samples which need to be tested/predicted.

(2) The process of evaluating a model

- Comparing the formulas of R^2 and $RMSE$, we can find that since R^2 is the ratio, it does not need to consider the range of its output value, avoiding the disadvantage of $RMSE$, and it is more convenient and reliable to use. Therefore, when comparing the quality of the models, I compare the scores of R^2first.
- When the model passed the first step: R^2evaluation, it is necessary to check its $RMSE$. The $RMSE$ of the model is expected to be as small as possible. But when the $RMSE$ at this time is greater than the $RMSE$ under other models of the same type (type: LR or LOR or SVR), how to decide whether to keep the model becomes a problem. The solution given here is that the $RMSE$ of the model to be tested

cannot be greater than the average RMSE of other models of the same type. Concretely, the method is to check whether the RMSE of the model to be tested is greater than one-fifth of the average RMSE. For instance, if the RMSE of the model to be evaluated is approximately 0.30, the mean of the RMSE is around 0.28. Since the difference between them is 0.02, less than 0.28 by one-fifth of the value, the model is given the opportunity to be further tested.

- Simply comparing R^2 and RMSE is not enough. The quality of the model is whether it properly avoids underfitting and overfitting or not. If the model has a poor fit in the validation set and the training set, suggesting that the model has undergone underfitting, then the model should be discarded directly. If the model performs much better or worse on the validation set than on the test set, then the model is not ideal because the former indicates that the model has strong randomness and the latter implies overfitting of the model. The target model to be searched has good and far-reaching performance on both the verification set and the test set. In this emotion prediction problem, if there R^2differs by more than 0.05, the model cannot be regarded as a good model.
- After finishing the above three-steps assessment, the last checking is simply executed. Namely, the size of training dataset comparison and the number of features comparison. A smaller size of the training dataset is preferred because it is more efficient, and the fewer number of features is expected for the same reason.

3.4.3.2 The process of testing model

The process of testing a model is closely related to if the model will be rationally estimated. The procedure of testing and selecting the optimal model is raised as an illustration in figure 3-11 and the following textual explanation:

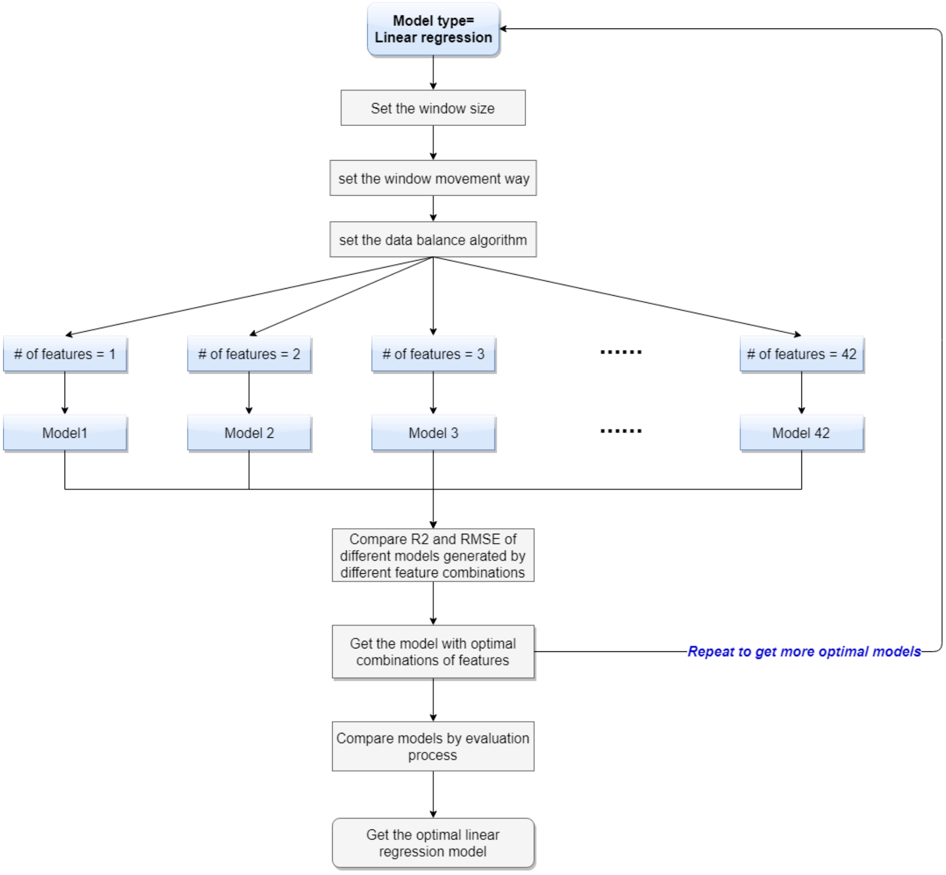
Model type=
Linear regression
Set the window size
set the window movement way
set the data balance algorithm
# of features = 1
# of features = 2
# of features = 3
......
# of features = 42
Model1
Model 2
Model 3
......
Model 42
Compare R2 and RMSE of different models generated by different feature combinations
Get the model with optimal combinations of features
Repeat to get more optimal models
Compare models by evaluation process
Get the optimal linear regression model

Figure 3-11 The procedure of testing and selecting the optimal model

According to the process of feature selection to choose the optimal feature combination:

i. Set 4 parameters, including the sliding window size, the movement way of the window, the data balance algorithm. and the model type.
ii. Compare the R^2 and RMSE of the different trained models generated by using a different number of features.
iii. Record the scores and the number of features of the model which has the best scores.
iv. In the same condition of model type, repeat i, ii, and iii to get many models.
v. preserve the scores and the related paraments (window size, movement way, data balance algorithm and the number of features) of the model A which owns the highest R^2 and RMSE.
vi. Repeat i, ii, iii, iv, and v, but this time the model is named as B.

vii. Repeat i, ii, iii, iv, and v, this time the model is named as C.

Model A, B and C are three models in 3 different case of model type, representing the optimal LR model, SVR model and LOR model respectively.

4 Results Evaluation and Discussion

There are 4 main parts in this chapter. Firstly, what experiments are conducted is described. Next, the optimal LR, SVR and LOR models are selected, analysed and compared. Thirdly, other discussions about the relationship between parameters and scores are given. Finally, the limitations of this research work are proposed.

4.1 Experiment description

Different parameter value settings lead to various experiments. The parameters that need to be set manually are the sliding window size, the window movement way and the model type. The combination of features as the fourth parameter which also affects the performance of the model, but it is automatically selected when the first three parameters mentioned above are fixed. Therefore, the number of trails in this project decided by the different combinations of window size, window movement way, and the model type.

The selection of these three parameters used for experiments is in the following range：

1) The size of the sliding window includes: 5, 10, 20
2) There are 4 types of the movement way of the sliding window
    - The separation of two adjacent sliding windows is one unit.
    - The sliding window moves forward approximately 1/3 of its size every time (i.e. if the sliding window size is 10, and the first position is in 0-9, then the second position of the window is beginning from 3).
    - The sliding window moves forward 1/2 of its size every time.

- The sliding window moves forward as much as its size every time.
- It is noticed that not all combinations of parameters within the above ranges have been tried. Moreover, considering big time-consuming in training SVR model, those potential experiments which take more than one minute to train a model is quitted, for instance, training an SVR model when the window size is 5.

The specific experiments are listed in the form of a table in Appendix 2.

3) The types of the model include linear regression, support vector regression and logistic regression.

4.2 Results demonstration and analysis

4.2.1 The optimal linear regression model

4.2.1.1 The description of the optimal linear regression model

1) The specific values of parameters being used to train the model

After comparing different LR models based on various parameters, the model with the following parameters is selected as the optimal LR model. The scores of the model are also listed after the parameters. It demonstrates 2 types of scores (R^2 and RMSE) used in two conditions. “R^2 for 10-fold cross validation” and “RMSE for 10-fold cross validation” are average scores when conducting 10-fold cross-validation. In the process of 10-fold CV, there are 10 specific models represented by 10 different linear equations being generated. Which one can stand for the optimal model to be evaluated by predicting the test data set is a key issue. The method I used borrowed the idea of a 10-fold CV which uses the average score as the final score. Similarly, each produced model is used for the test set prediction, and then the mean of the 10 scores is the final score which is used to measure the performance of the optimal linear model. Thus, "R^2 for the test dataset " and “MRSE for the test dataset” are the average scores.

Parameters:

- Window size = 20
- Movement way = 1
- Algorithm = 4
- Features = 41

Scores:

- R^2 for 10-fold cross validation = 0.521
- R^2 for the test dataset = 0.504
- RMSE for 10-fold cross validation = 0.246
- RMSE for the test dataset = 0.246

In this model, only the second feature is not been selected. The mask is visualised in the figure 4-1

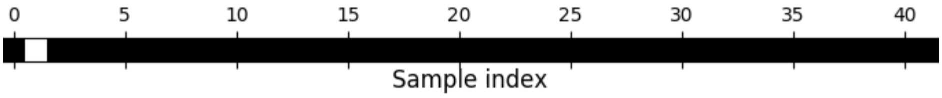

Figure 4-1 Simple index

2) The change of $\mathrm{R}^2$and RMSE (scores) of the model in 10-fold cross-validation in this model

As it mentioned before 10-fold CV caused 10-times data splitting and 10-times specific models being trained. The change of 10 R^2 and RMSE for the validation set and test set are shown as the figure 4-2.

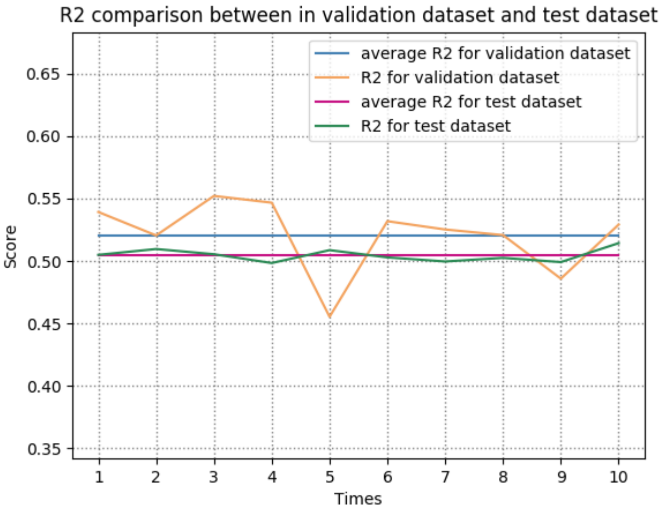

Figure 4-2 R^2 comparison between in validation dataset and test dataset

Comparing the blue and the pink line, showing that the average R^2 from the10-fold CV is higher than the score obtained from the process of predicting the test dataset. It reveals that the trained model over optimise the real-world prediction.

The change of the R^2 is relatively stable, while R^2 for the validation dataset has more substantial fluctuations. The most likely cause of this phenomenon is that the test set is fixed (segmented at the beginning), and the validation set changes in 10-fold cross-validation.

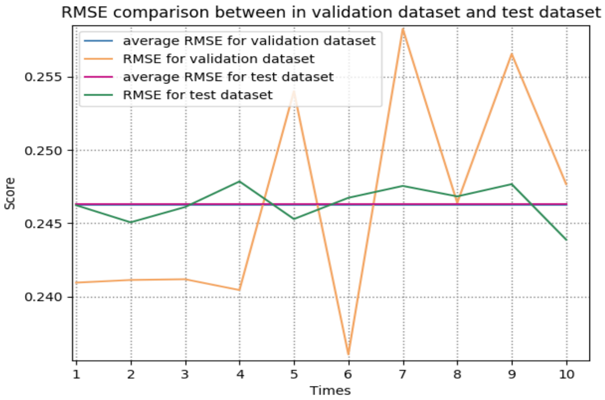

Figure 4-3 RMSE comparison between in validation dataset and test dataset

Different from the figure 4-2， figure4-3 display the tendency of RMSE, the blue and pink lines are almost coincident, and RMSE for validation dataset is extremely slightly higher than it in the test set. The oscillation amplitudes of these two RMSE has a similar trend with R^2 in the last line chart. Observing these two figures, it implies there is no absolute relationship between R^2 and RMSE, meaning in different training conditions, highest R^2 not accompany with the lowest RMSE.

3) The change of R^2 and RMSE (scores) of the model in different feature combinations

There are 42 different combinations of the features, the values of scores in each-time change are displayed in figure 4-4

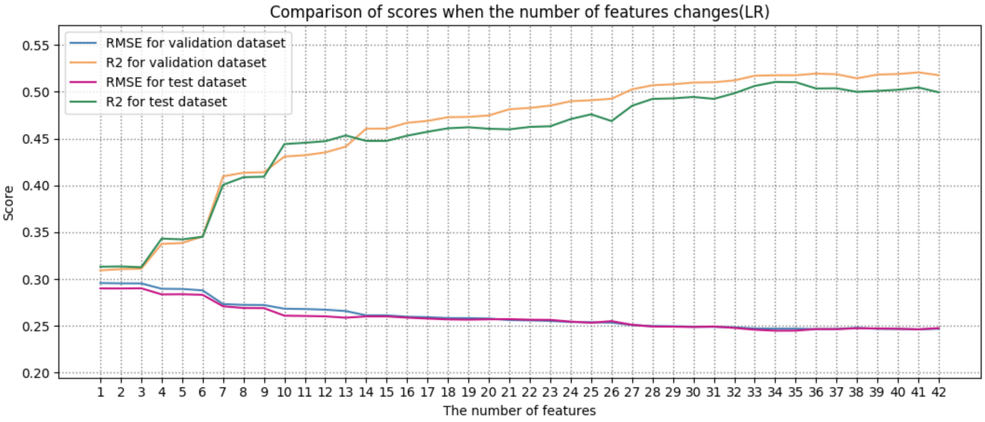

Figure 4-4 Comparison of scores when the number of features changes (LR)

As it is shown in the figure, the RMSE for both validation dataset and the test dataset gradually go down, in contrast, their R^2 increasingly rise until the number of features reaches to 41. The differences between the two R^2 scores slightly increase, which is opposite with RMSE score. This reminders people that the higher the degree of fit of the model, the greater the likelihood of over-fitting.

4.2.1.2 The analysis of the optimal support vector regression model

1) The comparison of scores of linear models in case of different window movement ways

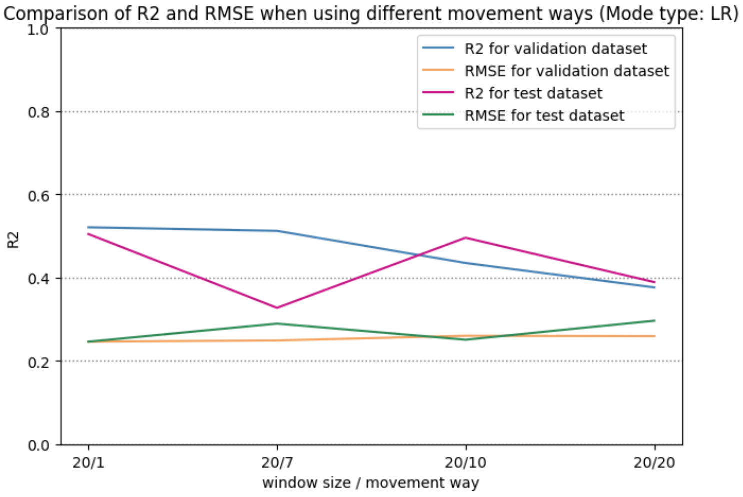

Figure 4-5 Comparison of R^2 and RMSE when using different movement ways (mode type: LR)

This figure 4-5 evidence that the synchronism of $\mathrm{R}^2$and RMSE changes under the same conditions. When the former rises, the latter decreases, and vice versa. In addition, when the movement way is 7, the trained model performs worst in the test dataset, and the gas between two R^2 scores are very big, which illustrates that the current model is not reliable.

2) The comparison of scores of linear models in case of different algorithms

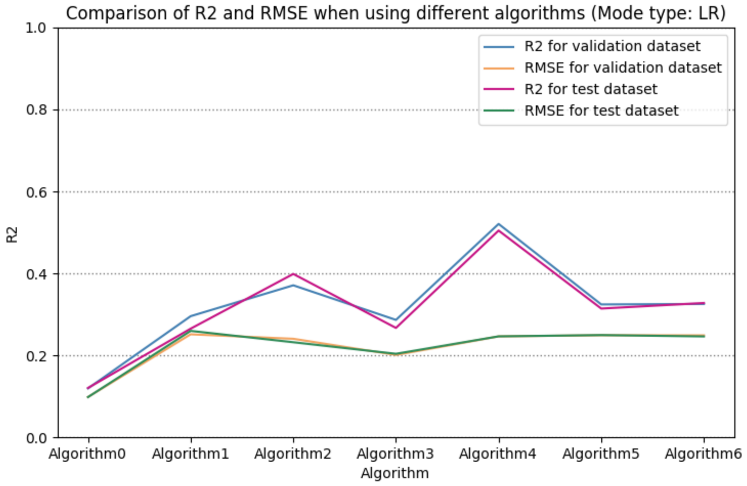

Figure 4-6 Comparison of R^2 and RMSE when using different algorithms (mode type: LR)

This graph visualises the performance of models based on different data balance algorithms. When there is no data balance algorithm being applied (i.e. algorithm0), the R^2 in any circumstances are worse, but the current RMSE is best. Algorithm1 also has a terrible performance, that is, with worst RMSE and the second poor R^2. Obviously, after trading off R^2 and RMSE, algorithm 4 is best in this phenomenon as it has highest R^2 and RMSE approximate to the mean.

4.2.2 The optimal support vector regression model

4.2.2.1 The description of optimal support vector regression model

1) The specific values of parameters being used to train the model

The parameters and the scores of the optimal SVR model are listed as follows:

Parameters:

- Window size = 20
- Movement way = 1
- Algorithms = 4
- Features = 40
- Kernel = “poly”

Scores:

- R^2 for 10-fold cross validation = 0.476.
- R^2 for the test dataset = 0.473.
- MRSE for 10-fold cross validation = 0.257.
- MRSE for the test dataset = 0.254.

This time there are 40 features being selected, being regarded as the best combination of features to make a higher accurate prediction. The feature selection result is visualised in figure 4-7. The black bars correspond to the selected features. The features that were eliminated were the second and 23rd.

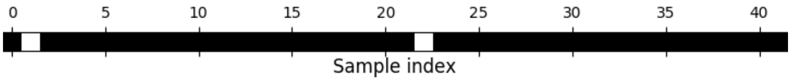

Figure 4-7 Sample index

2) The change of R^2 and RMSE (scores) of the model in 10-fold cross-validation in this model

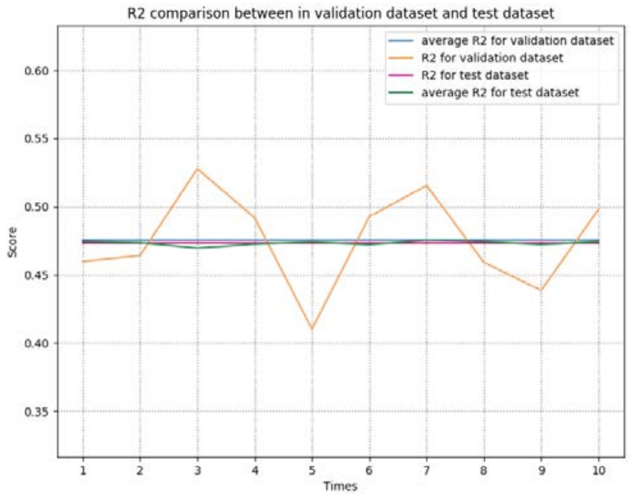

Figure 4-8 R^2 Comparison between in validation dataset and test dataset

The average scores of R^2 for test dataset and validation dataset are similar. R^2 obtained by testing the validation dataset has a visible fluctuation in the range of

0.42-0.58. Review the figure4-8 for linear regression, the predictive power of this SVR model is slightly worse but more stable.

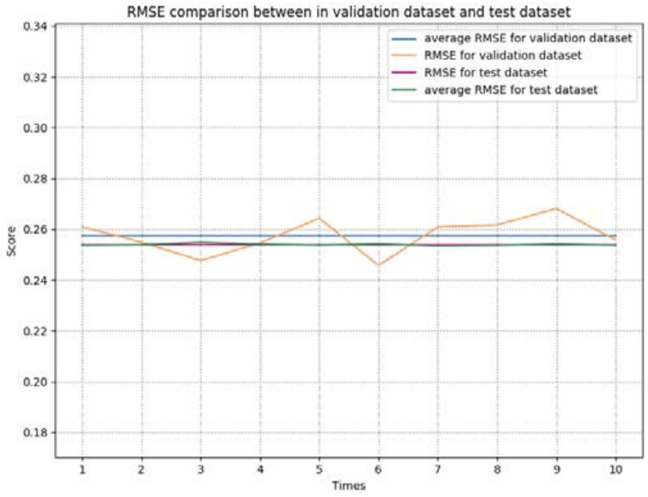

Figure 4-9 RMSE comparison between in validation dataset and test dataset

Average RMSE in 10-fold CV is a little higher than it in the test set. The fluctuation of RMSE is also obvious, but there are no big changes in RMSE for the test dataset.

3) The change of R^2 and RMSE (scores) of the model in different feature combinations

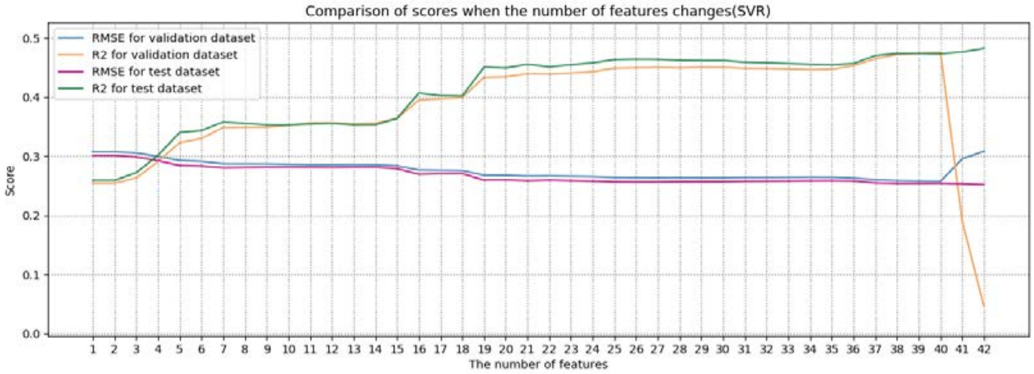

Figure 4-10 Comparison of scores when the number of features changes (SVR)

According to the figure 4-10, the tendency of the R^2 for validation dataset is continuously increase from 1 to 42, and its RMSE has the opposite change. However, for the test dataset, the appearance of the highest R^2 is when the number of features is 40. After that point, this score falls down dramatically, while RMSE also gets the lowest value at that point and has a visible increase after that.

4.2.2.2 The analysis of the optimal support vector regression model

4) The comparison of scores of linear models in case of different window movement ways

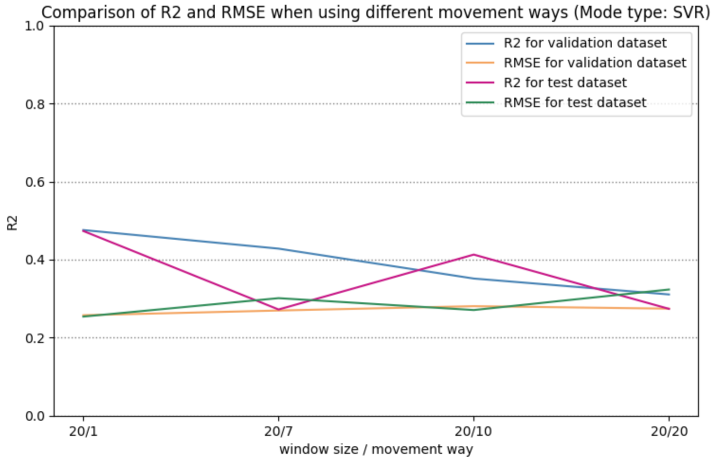

Figure 4-11 Comparison of R^2 and RMSE when using different movement ways (mode type: SVR)

Based on the figure 4-11, when the movement way of the window is 1, the performance of the trained model is best in the four cases. Specifically, R^2 for both validation and test dataset are highest, while the RMSE of them are the lowest. The worst performance of the model is when the movement way is as same size as the window (i.e. 20). At this moment, R^2 and RMSE for validation and test dataset are worst, opposition with the moment that the movement way is 1. When the movement way is 7 or 10, the performance of the model is unstable, namely the difference between R^2 for the validation dataset and it for the test dataset is big, especially in the case of movement way being 7.

5) The comparison of scores of linear models in case of different algorithms

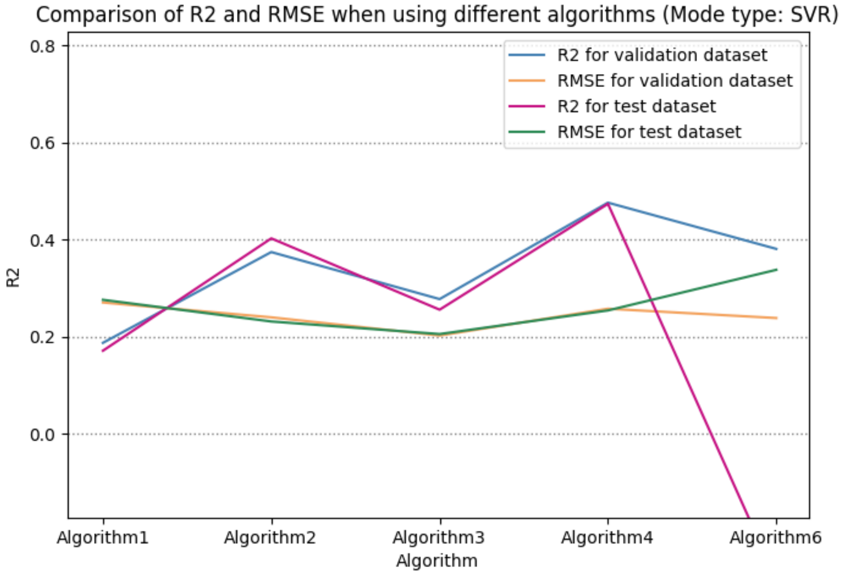

Figure 4-12 Comparison of *R^2* and RMSE when using different algorithms (mode type: SVR)

Observing the Algorithm-axis of the figure, there algorithm0 and algorithm5 are absent. Algorithm 0 is training the dataset without using any data balance methods, and algorithm 5 is the fifth data balance method. To train the SVR model, these two algorithms spend many few hours without results. Therefore, these two algorithms are abandoned temporarily. Through comparing these two types of scores, there is no doubt that algorithm4 perform better than others. There is an abnormal R^2 occurring in the case of the algorithm being 6. The value of the R^2 for the test set is below than 0, which is out of the normal range of R^2. It can be simply understood as that the model is too bad to run properly, and the concrete explanation of this phenomenon is given in the introduction of R^2 at the start of this chapter.

4.2.3 The optimal logistic regression model

4.2.3.1 The description of the optimal logistic regression model

1) The specific values of parameters being used to train the model

The optimal LR model has the following parameters and scores:

Parameters:

- Window size = 20
- Movement way = 1
- Algorithm = 4
- Features = 41

Scores:

- R^2 for 10-fold cross validation = 0.559
- R^2 for the test dataset = 0.555
- MRSE for 10-fold cross validation = 11.240
- MRSE for the test dataset = 11.155

The number of features being selected is also 41 and with the elements same as the optimal LR model. Hence, the second feature is not included in these 41 features.

2) The change of R^2 and RMSE (scores) of the model in 10-fold cross-validation in this model

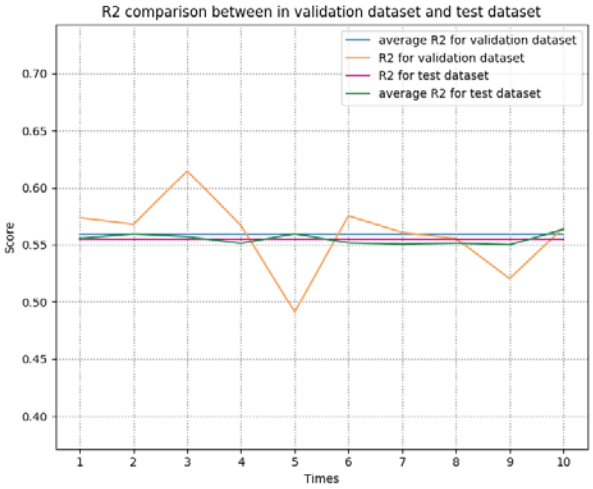

Figure 4-13 R^2 Comparison between in validation dataset and test dataset

Based on the figure, the two average R^2 are over 0.56, and the average R^2 for 10-fold CV is slightly higher than the other. Each R^2 for the test dataset is approximately equal to its mean. The range of R^2 for the validation dataset is between 0.50 and 0.62, implying the fluctuation still significant.

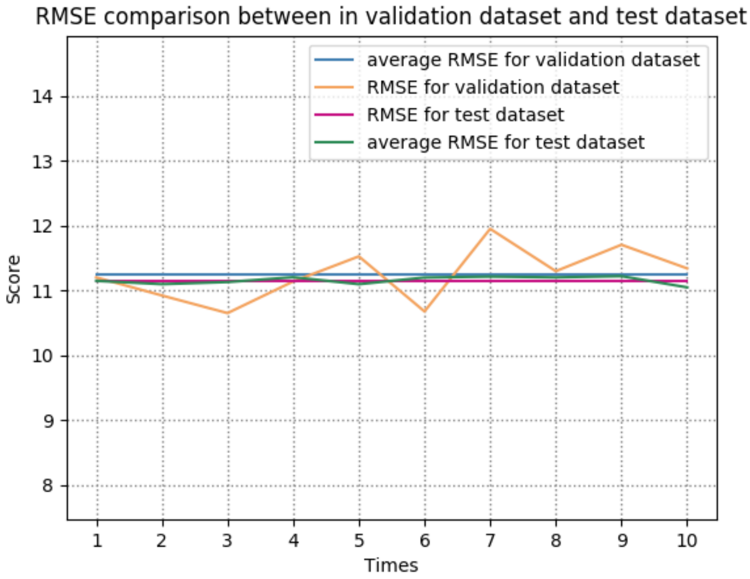

Figure 4-14 RMSE comparison between in validation dataset and test dataset

Different from the LR and SVR model, the RSME is distributed around 11, much larger than the RMSE got form the first two types of model. It does not mean the model is disappointing compared with LR models and SVR models because, after the novel transform of the LOR, the value of the output to be predicted is in the range of negative infinity to positive infinity. Compare figure4-13 and figure4-14, the tendency of RMSE is opposite to the tendency of R^2.

3) The change of R^2 and RMSE(scores) of the model in different feature combinations

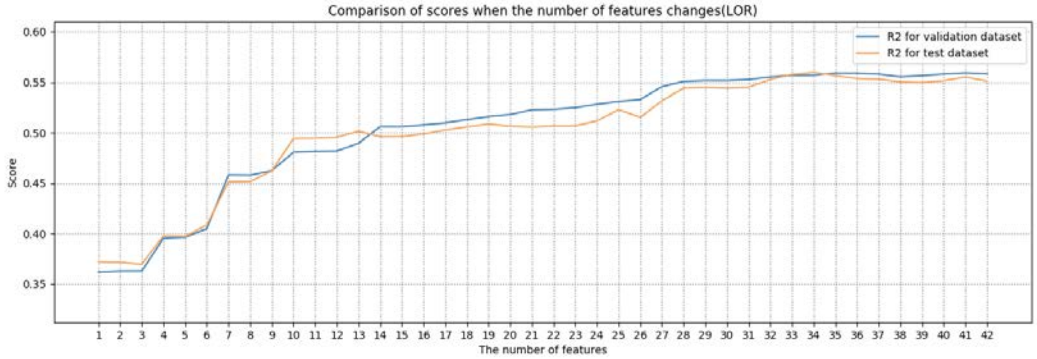

Figure 4-15 Comparison of scores when the number of features changes (LOR)

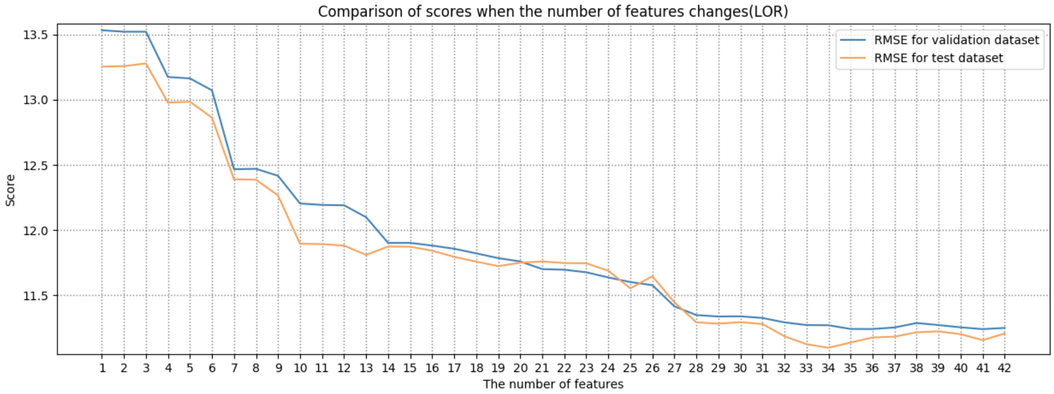

Figure 4-16 Comparison of scores when the number of features changes (LOR)

4) Figure4-15 and figure4-16 show that R^2 grows up, and the $RMSE$ decreases with the increase of the number of features until arriving at 41, indicating when the size and the movement way of the window are 40 and 1 respectively, and the algorithm4 is used as data balance method, the larger size of the features the better the model is.

4.2.3.2 Analysis of the optimal logistic regression model

5) The comparison of scores of linear models in case of different window movement ways

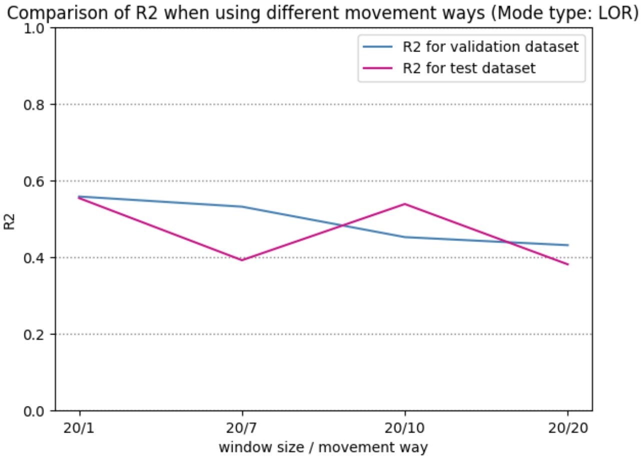

Figure 4-17 Comparison of R^2 when using different movement ways (mode type: LOR)

With the increase of the value of the movement way, the prediction ability of the model gradually decreases since the decline of R^2 showing in the figure. The worse model occurs when the movement is 20 because two R^2 are the lowest.

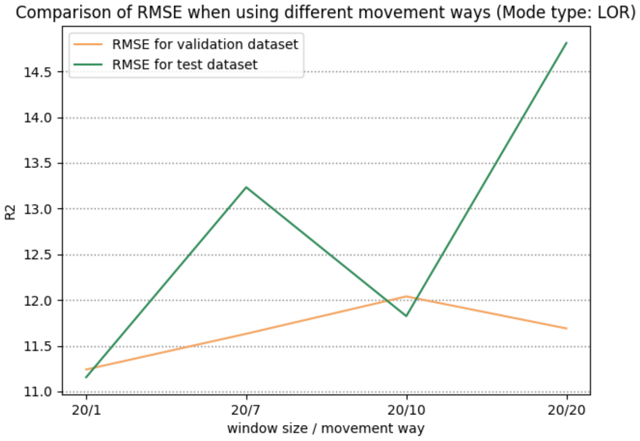

Figure 4-18 Comparison of RMSE when using different movement ways(mode type: LOR)

According to the figure 4-18, when the movement way is 20, both two RMSE are higher than in other conditions, and the distance of these two RMSE is biggest, meaning the model is unreliable. Combine the description of figure and figure, it deduces that simultaneously making 20 as the movement way and the size of the window is a worse choice.

6) The comparison of scores of linear models in case of different algorithms

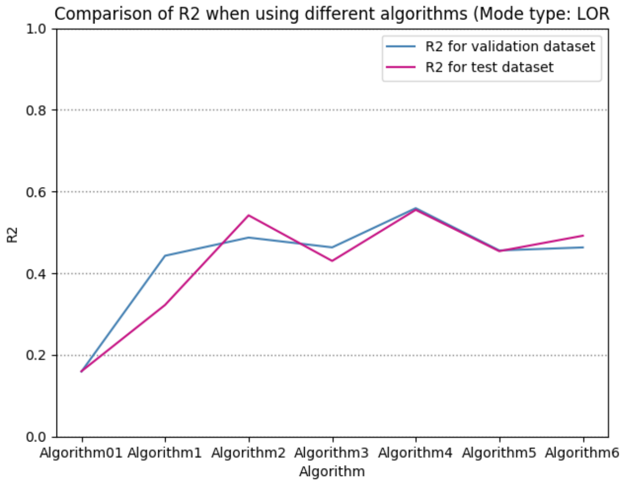

Figure 4-19 Comparison of R^2 when using different algorithms(mode type: LOR)

According to figure 4-19, in algorithm4, R^2 for validation dataset and test dataset are both highest. Using algorithm 5 also show a good performance in R^2.

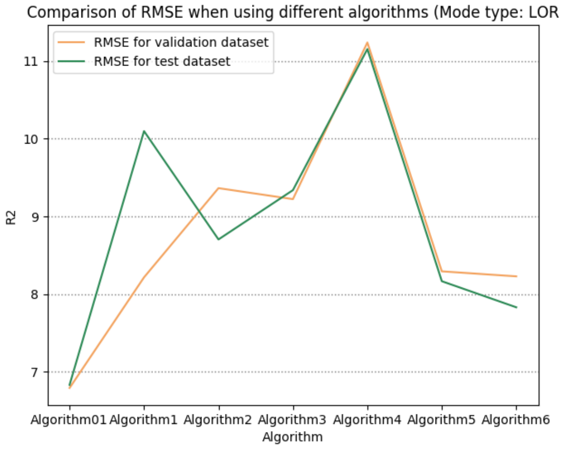

Figure 4-20 Comparison of RMSE when using different algorithms (mode type: LOR)

Figure 4-20 indicates that although the best R^2 is obtained using Algorithm 4, the corresponding cost is to sacrifice part of the RMSE score.

7) The comparison of scores of linear models

4.2.4 The final optimal model among LR, SVR and LOR

4.2.4.1 The comparison of scores among 3 different models

Every type of optimal model is evaluated and analysed as above. To compare them, the information of them are listed in the table as follows:

Table 4-1 Scores Comparison

	$\mathbf{RMSE}$ for validation dataset	$\mathbf{R}^2$ for validation dataset	$\mathbf{RMSE}$ for test dataset	$\mathbf{R}^2$ for test dataset
LR	0.246	0.521	0.246	0.504
SVR	0.257	0.476	0.254	0.473
LOR	11.240	0.559	11.155	0.555

Through checking the form, it indicates LOR has the highest training data fitting ability and test data prediction performance. Here since the scale of RMSE of LOR differs from the other two types of model, this criterion is not considered. The second-best model is LR, and the last is SVR. Moreover, in the experiment of training model, SVR takes much longer time to train a model than others because of the high complexity. Eventually, the final optimal model is LOR with other parameters that the window size is 20, movement way is 1, the data balance algorithm is 4, and the number of features is 41. In addition, this project also extracts this optimal combination of parameters: window size (20), movement way (1) and algorithm (4) in general.

4.2.4.2 The display of the result of the final optimal model

The following figure displays the result of predicting the test data.

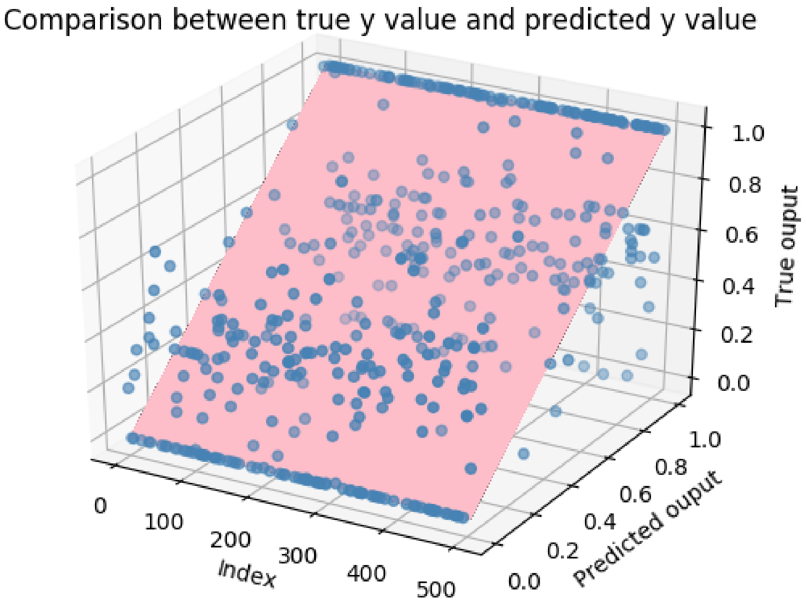

Figure 4-21 Comparison between the true y value and the predicted v value (from the front)

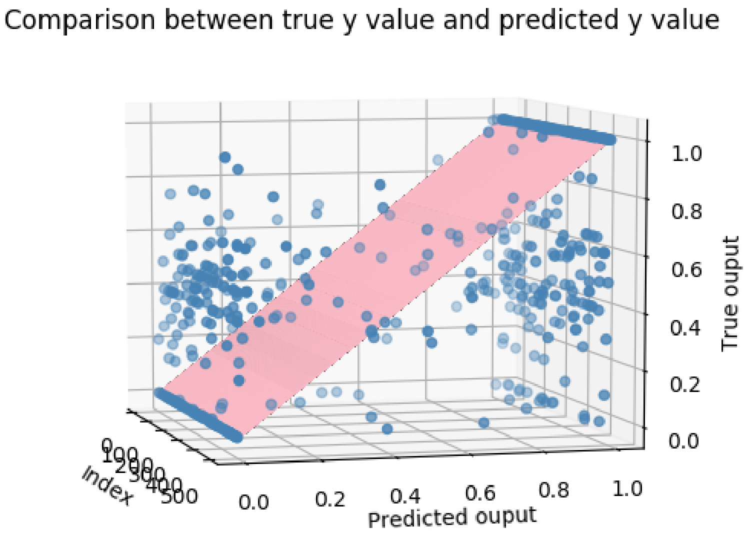

Figure 4-22 Comparison between the true y value and the predicted v value (from the side)

Comparing figure 4-21 and figure 4-22, the display of the result shows that when the actual output is 0 or 1, the model can give a good prediction. Relatively speaking, the model is not very good at predicting the value between 0 and 1. The model is prone to over-value or under-value the result.

4.3 The comparison between different parameters and scores

4.3.1 Comparison distribution of R^2 based on different models

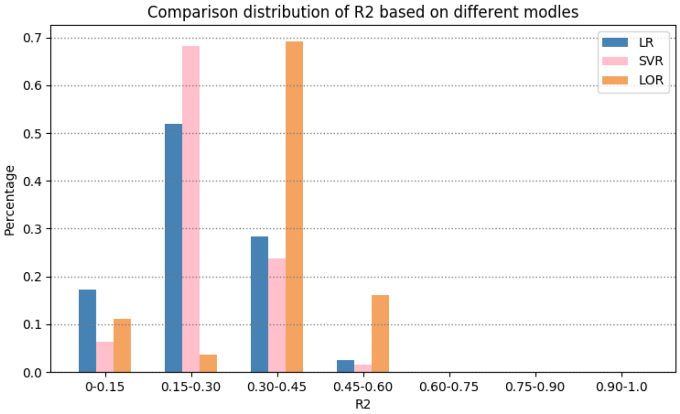

Figure 4-23 Comparison distribution of R^2 based on different models

In figure4-23, R^2 is the score when testing the validation dataset in 10-fold CV. In LR models, the biggest percentage of them have R^2 in range 0.15-0.3, the second percentage, second biggest but still much percentage of R^2 are concentrate on 0.30-0.45. Very few R^2 appear between 0.45 and 0.60.

R^2 in SVR has a similar pattern as in LR, but the gap between biggest and second biggest percentage enlarges. However, it is noticed that the number of SVR models is much smaller than LR models because many attempts to train models by using data sets processed by Algorithm 0 and Algorithm 5 failed. Therefore, the display of the distribution of R^2 in the figure is less reliable.

In LOR, most R^2 gather in 0.30-0.45, lesser of them in 0.45-0.60, and the smallest part of them is in 0.15-0.30.

Generally, the fitting ability of LOR-type models is better than the others.

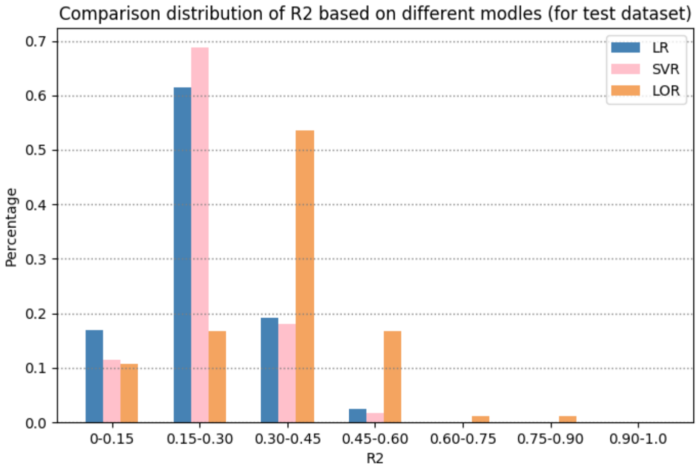

Figure 4-24 Comparison distribution of R^2 based on different models (for test dataset)

In figure 4-24, R^2 is the score when testing unknown real-world data instituted by the test dataset. This figure shows the prediction and generalisation performance of different type models.

Comparing the distribution changes of each model, we can find that the score distribution of LR and SVR shifts from right to left. In particular, for LR the ratio of original R in the range of 0.3-.04 is reduced, and loss proportions contribute to 0.15-0.30. Similarly, in SVR the partial percentage in 0. 3-0. 45 is converted to 0-0. 15, and the magnitude of the conversion is relatively small compared to LR. The partial percentage of the LOR in the range of 0.30-0.45 is transferred to the range on the left and right sides. It can be seen that, except for LOR, the prediction ability for the test set is generally lower than the prediction capability for the verification set.

Overall, these two figures demonstrate that the LOR is the optimal model among these three types of models in this project.

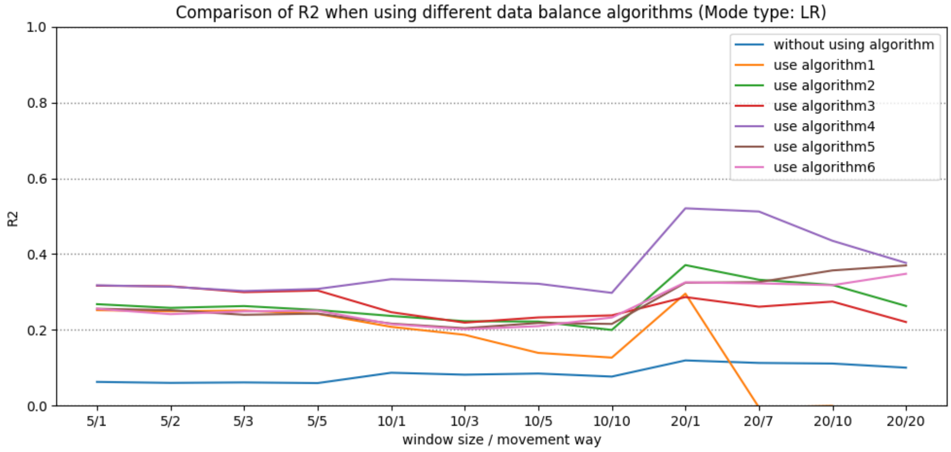

Figure 4-25 Comparison of R^2 when using different data balance algorithms (mode type: LR)

In figure 4-25, R^2 is obtained by testing the validation dataset in 10-fold CV. Whatever the window size or the movement way is, using algorithm4 can get a higher score and provide a more accurate prediction.

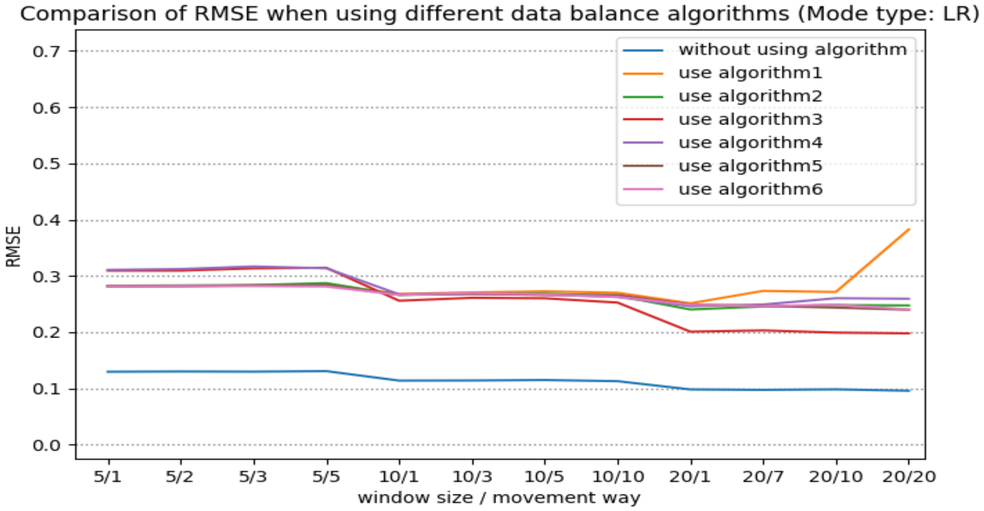

Figure 4-26 Comparison of RMSE when using different data balance algorithms (mode type: LR)

In figure 4-26, RMSE is obtained by testing the validation dataset in 10-fold CV. Without using any data balance method can make the RMSE be lowest, but according to the model's R^2, its prediction performance is extremely poor. Thus, balancing data is necessary even if it is at expense of sacrificing RMSE scores.

Then comparing RMSE of different algorithms, it is found that the algorithms4 is at a medium level. Therefore, algorithm4 is more suitable for training LR model.

4.4 Limitations

4.4.1 The limitation from not being able to try all possible parameters.

In experiments, not all possible window sizes and movement ways have been tried. Possible attempts are the result of "(2 * 2+ 3 *3 +4 * 4... +1000* 1000) *7*3". This because the sliding window size can be an arbitrary number from 2 to 1000, 999 kinds in total. In other words, it is the value removing 1 from the number of sample sized chunks in each given data file. For each window, there are n types of movement way, of which n is the same value as its own size. Plus, 7 is 1(do experiment without using data balance algorithm) plus 6 (6 types of data balance algorithms). Additionally, 3 is the number of ML model types used in this project.

Since the number of potential experiments is too large, it is not feasible to manually try one by one. If expecting to achieve it by writing a program, it requires high performance on the computer because of its large amount of computation. It cannot be completed by ordinary computers.

4.4.2 The limitation from the way of how to obtain the math expression of the final optimal model

The final optimal model is selected through comparing scores of three optimal models (LR, SVR, and LOR model). The scores representing the final optimal is originated from averaging 10 groups of scores generated by 10-fold CV, therefore, there are 10 mathematical expressions for this model. The foremost issue is which expression can be used to represent the model. This is important because this mathematical expression is used to predict the test set to measure the generalisation performance of the model in the real world.

From my view, there are three reasonable solutions:

1) Choose the mathematical expression which is most close to the average scores to act on behalf of the final optimal model.
2) Use the dataset that integrates training and validation set as a new training dataset to train a model and gain a mathematical expression based on the same 4 parameters (i.e. the size and the movement way of sliding window, the combination of features and the model type)
3) All 10 mathematical expressions of the model are used to predict the test dataset 10 times. Then the mean score of 10-times predictions is calculated and used to measure and assess the performance of the model.

In this project, the third method is applied. The third method is inspired by using CV to calculate the average score. It is relatively accurate, but time-consuming. In the practical world, the first and second solutions may be more proper.

4.4.3 The limitation from the way of how to evaluate a model.

It is a better way to assess a model from as many aspects as possible. This paper evaluates a model mainly based on R^2 and RMSE of the model and the ease of implementation of the model. More model evaluation methods are expected to be applied to have a more accurate and reasonable evaluation of the model.

4.4.4 The limitation from the fix test dataset.

The test data set is the first 10 per cent data in the updated dataset, and the remaining data set is used to do 10-fold CV. Since the test data set is fixed and not very big, it is not accurate enough to evaluate the model's ability to adapt to the world based solely on this part of the data. It is suggested that the principle of CV can be used. The first time is taken as the first 10% as the test set, the remaining part is used to perform the ten-fold CV. Then in the second time, other 10% data is extracted as the test set, and the remaining part is used in 10-fold CV, and so on.

5 Conclusion and future work

To achieve the aims “emotion detection and percentage prediction”, many tasks have been completed, including:

- Propose 6 data balance algorithms to improve the robustness of the model.
- Automatically select the proper feature combinations which can maximum the score of the model.
- Determine the optimal linear regression, logistic regression, support vector regression model and the final optimal model by the self-defined evaluation process.
- For each model produced by 10-fold CV, the prediction and generalisation performance on the test set was evaluated for 10 times instead of the test by once, which provides a more proper evaluation method.

After analysing the result of hundreds of experiments, the conclusion is drawn as follows:

- When the optimal parameters for the window size, the window movement way, and algorithm are 20,1,4 respectively. The optimal LR, LOR and SVR are all based on these parameters
- Logistic regression performs best in this project. Generally, in the same condition, the scores of the LOR models higher than the other two types of models.
- The complexity of the support vector regression model usually higher, and it takes a longer time to train a model. Even so, the scores of the model are still the lowest, so it is the least recommended.
- Data balance algorithm 4 usually performs best in all algorithm. Algorithm1 often show the worst result because the size of the dataset after using this algorithm is greatly affected by the size and distribution of the dataset.

The future work about this thesis is suggested as follows:

- Propose more rational data balance methods to improve the prediction performance of the model
- Test and assess the performance of the model by comparing with the human performance
- Combine the ML model with audio and video processing to improve the robustness of the model.

6 Appendix A How to run the programme

This program is implemented in Python3.

1. If you want to train a model, please run “models.py”. In the user interface, it will give you some instructions, please follow them.
2. Some results are put in the file “resultVisualisation_1”, “resultVisualisation_2”, “resultVisualisation_3”. You can run them if you like.
3. In the file “resultVisualisation_2”, the CSV file store all experiments result.
4. The dissertation is put in the file which is named as “report”.

7 Reference

[1] L. Barracliffe, O. Arandjelovic, and G. Humphris, *A Pilot Study of Breast Cancer Patients: Can Machine Learning Predict Healthcare Professionals' Responses to Patient Emotions?* 2017.

[2] T. E. Joiner Jr *et al.*, "Can positive emotion influence problem-solving attitudes among suicidal adults?," *Professional Psychology: Research and Practice,* vol. 32, no. 5, p. 507, 2001.

[3] J. G. Rabkin, R. Remien, J. B. Williams, and L. Katoff, "Resilience in adversity among long-term survivors of AIDS," *Psychiatric Services,* vol. 44, no. 2, pp. 162-167, 1993.

[4] K. H. Dow, B. R. Ferrell, S. Leigh, J. Ly, and P. Gulasekaram, "An evaluation of the quality of life among long-term survivors of breast cancer," *Breast cancer research and treatment,* vol. 39, no. 3, pp. 261-273, 1996.

[5] P. L. Morris, R. G. Robinson, and J. Samuels, "Depression, introversion and mortality following stroke," *Australian & New Zealand Journal of Psychiatry,* vol. 27, no. 3, pp. 443-449, 1993.

[6] C. Villemure and C. M. Bushnell, "Cognitive modulation of pain: how do attention and emotion influence pain processing?," *Pain,* vol. 95, no. 3, pp. 195-199, 2002.

[7] M. de Wied and M. N. Verbaten, "Affective pictures processing, attention, and pain tolerance," *Pain,* vol. 90, no. 1-2, pp. 163-172, 2001.

[8] D. T. D., M. T. V., and S. J. F., "Patient treatment preferences in localized prostate carcinoma: The influence of emotion, misconception, and anecdote," *Cancer,* vol. 107, no. 3, pp. 620-630, 2006.

[9] E. Wittkower, "Studies on the Influence of Emotions on the Functions of the Organs: (Including Observations in Normals and Neurotics)," *Journal of Mental Science,* vol. 81, no. 334, pp. 533-682, 1935.

[10] A. J. Mitchell, D. W. Ferguson, J. Gill, J. Paul, and P. Symonds, "Depression and anxiety in long-term cancer survivors compared with spouses and healthy controls: a systematic review and meta-analysis," *The Lancet Oncology,* vol. 14, no. 8, pp. 721-732, 2013/07/01/ 2013.

[11] C. Zimmermann *et al.*, "Coding patient emotional cues and concerns in medical consultations: the Verona coding definitions of emotional sequences (VR-CoDES)," *Patient education and counseling,* vol. 82, no. 2, pp. 141-148, 2011.

[12] J. Turner and B. Kelly, "Emotional dimensions of chronic disease," *Western Journal of Medicine,* vol. 172, no. 2, p. 124, 2000.

[13] "Living with and treating rare diseases: Experiences of patients and professional health care providers," *Qualitative health research,* vol. 25, p. 636, 2015.

[14] R. M. Epstein, ed, 2007.

[15] A. L. Suchman, K. Markakis, H. B. Beckman, and R. Frankel, "A model of empathic communication in the medical interview," *Jama,* vol. 277, no. 8, pp. 678-682, 1997.

[16] W. Levinson, R. Gorawara-Bhat, and J. Lamb, "A study of patient clues and physician responses in primary care and surgical settings," *Jama,* vol. 284, no. 8, pp. 1021-1027, 2000.

[17] P. Butow, R. Brown, S. Cogar, M. Tattersall, and S. Dunn, "Oncologists' reactions to cancer patients' verbal cues," *Psycho-Oncology,* vol. 11, no. 1, pp. 47-58, 2002.

[18] C. Heaven, P. Maguire, and C. Green, "A patient-centred approach to defining and assessing interviewing competency," *Epidemiology and Psychiatric Sciences,* vol. 12, no. 2, pp. 86-91, 2003.

[19] L. Del Piccolo, H. De Haes, C. Heaven, J. Jansen, W. Verheul, and A. Finset, "Coding of health provider talk related to cues and concerns," ed: Manual, 2009.

[20] L. Del Piccolo, A. Finset, and C. Zimmermann, "Verona coding definitions of emotional sequences (VR-CoDES). Cues and concerns manual; 2009," *European Association for Communication in Healthcare (EACH),* 2009.

[21] A. Wright, G. Humphris, K. L. Wanyonyi, and R. Freeman, "Using the verona coding definitions of emotional sequences (VR-CoDES) and health provider responses (VR-CoDES-P) in the dental context," *Patient education and counseling,* vol. 89, no. 1, pp. 205-208, 2012.

[22] F. Dicé, P. Dolce, and M. F. Freda, "Exploring emotions and the shared decision-making process in pediatric primary care," *Mediterranean journal of clinical psychology,* vol. 4, no. 3, 2016.

[23] A. Finset, L. Heyn, and C. Ruland, "Patterns in clinicians' responses to patient emotion in cancer care," *Patient education and counseling,* vol. 93, no. 1, pp. 80-85, 2013.

[24] Y. Zhou *et al.*, "Applying the Verona coding definitions of emotional sequences (VR-CoDES) in the dental context involving patients with complex communication needs: An exploratory study," *Patient education and counseling,* vol. 97, no. 2, pp. 180-187, 2014.

[25] V. Ramalingam, A. Pandian, A. Jaiswal, and N. Bhatia, "Emotion detection from text," in *Journal of Physics: Conference Series*, 2018, vol. 1000, no. 1, p. 012027: IOP Publishing.

[26] J. F. Cohn and G. S. Katz, "Bimodal expression of emotion by face and voice," in *Proceedings of the sixth ACM international conference on Multimedia: Face/gesture recognition and their applications*, 1998, pp. 41-44: ACM.

[27] T. Nguyen, I. Bass, M. Li, and I. K. Sethi, "Investigation of combining SVM and decision tree for emotion classification," in *Multimedia, Seventh IEEE International Symposium on*, 2005, p. 5 pp.: IEEE.

[28] Y. Torao, S. Naruki, Y. Kaori, and N. Masahiro, "An emotion processing system based on fuzzy inference and subjective observations," *Information sciences,* vol. 101, no. 3-4, pp. 217-247, 1997.

[29] A. Seyeditabari, N. Tabari, and W. Zadrozny, "Emotion Detection in Text: a Review," *arXiv preprint arXiv:1806.00674,* 2018.

[30] L. Devillers, L. Vidrascu, and L. Lamel, "Challenges in real-life emotion annotation and machine learning based detection," *Neural Networks,* vol. 18, no. 4, pp. 407-422, 2005.

[31] R. Hirat and N. Mittal, "A Survey On Emotion Detection Techniques using Text in Blogposts," *International Bulletin of Mathematical Research,* vol. 2, no. 1, pp. 180-187, 2015.

[32] C. Strapparava and R. Mihalcea, "Learning to identify emotions in text," in *Proceedings of the 2008 ACM symposium on Applied computing*, 2008, pp. 1556-1560: ACM.

[33] P. Chiranjeevi, V. Gopalakrishnan, and P. Moogi, "Neutral face classification using personalized appearance models for fast and robust emotion detection," *IEEE Transactions on Image Processing,* vol. 24, no. 9, pp. 2701-2711, 2015.

[34] J. Python, "Python programming language," in *USENIX Annual Technical Conference*, 2007.

[35] G. van Rossum, "Python is an interpreted high-level programming language for general-purpose programming. Created by Guido van Rossum and first released in 1991, Python has a design philosophy that emphasizes code readability, notably using significant whitespace. It provides constructs that enable clear programming on both small and large scales.[27] Python features a dynamic type system and automatic memory management. It supports multiple programming paradigms, including object-oriented, imperative, functional and procedural, and has a large and comprehensive standard library.[28] Python interpreters are available for many operating systems. CPython, the reference implementation of Python, is open source software [29] and has a community-based development model, as do nearly all of its variant implementations. CPython is managed by the non-profit Python Software Foundation."

[36] M. Nosrati, "Python: An appropriate language for real world programming," *World Applied Programming,* vol. 1, no. 2, pp. 110-117, 2011.

[37] L. Prechelt, "An empirical comparison of c, c++, java, perl, python, rexx and tcl," *IEEE Computer,* vol. 33, no. 10, pp. 23-29, 2000.

[38] O. K. Tarald, "Cautionary note about R2," *The American Statistician,* vol. 39, no. 4 Pt 1, pp. 279-285, 1985.