|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | |  |  | | --- | --- | | For office use only | | | T1 | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | | T2 | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | | T3 | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | | T4 | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | | |  | | --- | | Team Control Number **8784** | |  | | Problem Chosen **A** | | |  |  | | --- | --- | | For office use only | | | F1 | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | | F2 | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | | F3 | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | | F4 | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | |   **2018 HiMCM Summary Sheet** |

Blablablablabla

**Key Words:** AHP, Principle Component Regression, Bayes Distinction, BP Neural Network Fitting, XG Boosting algorithm

**News Release**

Blablabla

**Contents**

1. **Background**
   1. **Research Background**
   2. **Restatement of the Problem**
   3. **Research Method and Train of Thinking**
2. **Assumptions**
   1. **Assumptions**
   2. **Definitions**
3. **Data Procurement and Process**
   1. **Data Cleaning and Interpolation**
   2. **Cluster**
   3. **Ideal Solution**
   4. **Principal Component Analysis**
4. **Modeling**
   1. **Basic Statistics**
   2. **AHP Method**
   3. **Linear Regression**
   4. **KNN Algorithm**
5. **Optimization**
   1. **Principal Component Regression**
   2. **Bayes Distinction**
   3. **BP Neural Network Fitting**
   4. **XG Boosting Algorithm**
6. **Sensitivity Analysis**
7. **Conclusion**
   1. **Strength and Weakness**
   2. **Comparison of the top 10 Roller Coasters**
   3. **Concept and design for a user-friendly app**
8. **Reference**
9. **Appendix**
   1. **MATLAB Code**
   2. **PYTHON Code**
   3. **Final Result**
10. **Background**
    1. **Research Background**

Roller coaster is an exciting entertainment fascinated by many of today's youngsters because of its stimulation and pleasure. However, the rating system of roller coasters is relatively scarce and mainly based on people's own experience, lacking quantitative analysis on roller coasters' different traits. For instance, Coaster Buzz posted the current top 100 roller coasters on its website, but the rating process is largely based on its members' track records and subjective experience inputs. Admittedly, one of the method's main advantages proves to be the vast sample size and the ratings' rigorous selection in order to exclude the anomalies. But even with the rating results refreshed weekly, the poll on the internet only reflects the opinion of any one person and the riders who provide their experience and scores may mostly come from the same region, so the ranking is not based on the world's scale and will certainly still lose some great rides. Therefore, the most rating methods are highly unstable and unconvincing, making the roller coaster-choosing process inaccurate and the riders being dissatisfied. According to the current needs and lack of quantitative methods of rating, a proper method for ranking the roller-coasters is in dire need.

* 1. **Restatement of the Problem**

The question is based on the fact that nowadays roller coaster ranking systems are largely dependent on riders' own subjective inputs, with few considering the roller coasters' own properties. Providing us with the basic information of 300 roller coasters around the world, the question asked us to decide the top ten roller coasters using quantitative assessing methods, compare them with other methods currently being used and analysis the strength and weakness. Besides, we are required to develop the concept of a user-friendly APP which aims to help the potential riders finding the proper roller coasters that will satisfy their needs. Finally, we write a News Release to publicize our quantitative methods, the result of top-10 roller coasters based on the data given and the concept of our newly-designed APP.

* 1. **Research Method and Train of Thinking**

We do the data cleaning first and interpolate the missing data, while extracting useful and relevant data and conducting basic statistics for further research. Next, we come to the data procurement part to examine whether the score of the roller coasters online can be a training set of our model with the help of Principal Component Analysis. Then we apply the results of data procurement for modeling. In the modeling process, we apply results from Principal Component Analysis to Analytical Hierarchy Process, KNN, and Linear Regression. At this point, we have reached the conclusion of the rank of different independent factors. Furthermore, we conduct optimization to each model. We optimize KNN by Bayes Distinction and Linear Regression by Principal Component Regression, while we optimize Entropy of Information to BP Neural Network Fitting. Afterward, we employ the XG Boosting algorithm to synthesize the three methods and reach the conclusion Finally, we compare our rating and raking with those online and design the notion of our desired application. Figure 1 below presents the whole modeling process.

Beginning

Data Cleaning and Interpolation

Ensure that all the data are valid

Examine the ranking made by us and the ranking online

XGBoosting Algorithm

Sensitivity Analysis

Conclusion

Cluster

PCA analysis

Data Examining

Ideal value method

BP Neural Network Fitting

Principal Component Regression

Optimization

Bayes Distinction

Linear Regression

AHP

Basic Statistics

KNN

Modeling

Synthesizing the three methods

Figure 1: The flow chart of the whole modeling process

1. **Assumptions**
   1. **Assumptions**

* All the data are credible and reliable, which means they have no error.
* The parameters of the roller coasters are constant and do not change at different time in a day. For instance, the speeds of the roller coasters are always the value given in the table. No matter it is in the morning or evening, or no matter how many people there are on the roller coasters, they will always travel at the speeds given.
* Blablabla
  1. **Definitions**

Table 1: the definition of notations

|  |  |
| --- | --- |
| Notation | Definition |
|  | The element in Row and Column in matrix |
|  | The independent variables matrix |
|  | Row vector of independent variables |
|  | Row vector of dependent variables |
|  | The algebra average of several data |
|  | The Mahalanobis distance of the data |
|  | The covariance matrix |
|  | The original variable |
|  | The New variable |
|  | The number of samples |
|  | The number of variables in each sample |
|  | The standardized data at row and column |
|  | The data at row and column before standardization |
|  | The correlation coefficient matrix in principal component analysis |
|  | The characteristic roots or eigenvalues in Weight determination Technique |
|  | The characteristic vectors |
|  | The th value of the characteristic vectors |
|  | Weight vector in AHP |
|  | The number of choices of target layer in AHP |
|  | The eigenvector in AHP |
|  | Coefficient matrixes of the original data |
|  | Coefficient matrixes of Principal Component Regression |
|  | The probability that satisfies condition |
|  | Reliability in Regression |
|  | Parameters to be estimated of the ensemble in Regression |
|  | The confidence upper limit in Regression |
|  | The confidence lower limit in Regression |
|  | Posteriori probability in Bayes Distinction |
|  | Priori probability in Bayes Distinction |
|  | The frequency at which the sample appears in Bayes Distinction |
|  | The ensemble in Bayes Distinction |
|  | Probability density function of in Bayes Distinction |
|  | The priori probability of In Bayes Distinction |
|  | The number of in Bayes Distinction |
|  | The conditional probability of wrongly categorizing the sample of to the ensemble |
|  | The loss caused by the wrong categorization |
|  | A division of a set of distinction samples |
|  | The average wrong distinction loss |
|  | The overall loss of each classifier |
|  | Classification function |
|  | function of each classifier to reduce the loss |
|  | The score of the data to show the accuracy of the prediction |

1. **Data Procurement and Process**
   1. **Data Cleaning and Interpolation**

Original Data

Speed (mph): 74.0

Duration (min:sec): 3:00

Name: Big One

Type: Sit Down

Drop (feet): 205.0

Construction: Steel

Park: Blackpool Pleasure Beach

Geographic Region: Europe

Status: Operating (discarded due to no variation)

Number of Inversions: 0

City/Region: Blackpool

Inversions (YES or NO): NO (discarded due to already contained)

Height (feet): 213.0

City/State/Region: Lancashire, England

Year/Date Opened: 1994

G Force: 3.5 (discarded due to too many missing data)

Country/Region: United Kingdom

Length (feet): 5497.0

Length (feet): 5497.0

Length (feet): 5497.0

Vertical Angle (degrees): 65 (discarded due to too many missing data)

Duration (sec): 180 (Interpolation if needed)

Number: 32 (added by given order)

Construction: 1

Geographic Region: 2

Height (feet): 213.0 (Interpolation if needed)

Speed (mph): 74.0 (Interpolation if needed)

Length (feet): 5497.0 (Interpolation if needed)

Length (feet): 5497.0

Length (feet): 5497.0

Drop (feet): 205.0 (Interpolation if needed)

Type: 1

Figure 2: Data Processing diagram

As we downloaded the data, we first number the 300 roller coasters from 1 to 300 and do the data cleaning as the foundation of the entire model. We remove the drop column, the G Force column and the Vertical Angle column since there are more than a half of the data missing, which renders it void for us to interpolate the missing value. We also remove the status column, since all the roller coasters are operating. Then with the help of XLRD and XLWR module in PYTHON, we convert the expression of the duration cells from both the minutes and seconds into seconds only. We also numerate the Geographic Region column, the Construction column and the Type column. For the Geographic Region column, we employ 1 to 8 represent Asia, Europe, North America, Central America, South America, Middle East, Oceana, and Russia respectively. For the Construction column, we use 1 to 2 represent steel and wood respectively. For the Type column, we use 1 to 6 represent sit down, inverted, stand up, suspended, flying, and wing respectively. We also notice that some of the Type cells are filled in steel or wood, which is not a possible choice of Type, which we use 0 to represent the two choice. We removed the unit in the cells of Height in order that it is able to be dealt with further.

We discover that some of the data in the Height, Speed, Length, and Duration column are missing, thus we consider that we use the interpolation method to fill in the missing number. We examine the correlation coefficients between the columns, and find that the correlation coefficient between Height and Speed is 0.836280084187907, and the one between Length and Duration column is 0.619704366781674, indicating that the two groups of column reveals a strong tendency of correlating, which means we can use the two columns in each group to interpolating the missing data of each other. We sort the interpolating variable and calculate the arithmetic means of the interpolated variable if an interpolating variable refers to more than one interpolated variable in the given data set before we utilize Piecewise Cubic Hermite Interpolation to interpolate our variable. Similarly, we do the same process for the rest 3 columns and fill in all the data.

The reason why Piecewise Cubic Hermite Interpolation is suitable for our problem is that it avoids the oscillation between the point series, while we do not pay much attention to the smoothness of the interpolation function. We eliminated some data that miss both interpolating variable and interpolated variable.

Finally, we obtain 6 variables that we mainly use, which are Geographic Region, Construction Type, Year Opened, Height, Speed, Length, Duration, and Number of Inversions. The columns that are not mentioned above are regarded as the identification of each roller coaster, which will not be used for modeling. The 293 data after cleaning can be seen in the appendix. The following figure 2 illustrates the process mentioned above.

* 1. **Cluster**

We would like to rank the roller coasters at the beginning and compare the ranking with a ranking of the scoring system online. If these two are similar, we can regard the online scoring system as a learning set and establish a model to rate all the roller coasters. [11]

We utilize cluster to decide which of the roller coasters are similar. It can be predicted that similar roller coasters are more likely to have the similar rating and ranking, therefore we can divide all the roller coasters into several groups. If the roller coasters that are in the same group are more likely to lie in the same online score interval, such as the high score or the low score, we can say that our ranking system is consistent with the online scoring system, which makes it viable for us to establish a model with the online system.

We use Mahalanobis distance for clustering and draw the dendrogram. The formula is as the following formula 1.



(1)

Among the formula, and denote two row vectors; denotes the covariance matrix; denotes the obtained Mahalanobis distance of the data. The result is shown in figure.

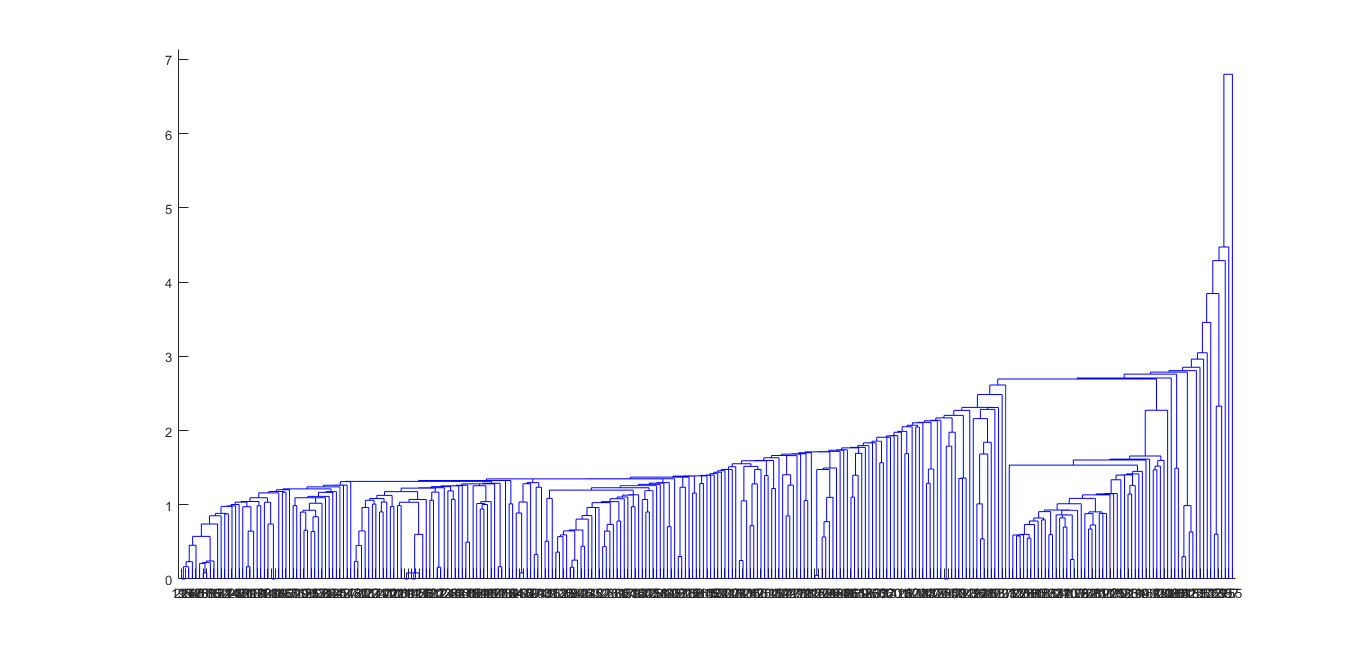


Figure 3: Mahalanobis Clustering Dendrogram, some of the categories contain too less roller coasters

From the figure 3 above, we can see that some of the categories contain too less roller coasters, which shows that this method is difficult to set the roller coasters apart. Hence, we consider to use other methods.

* 1. **Ideal Solution**

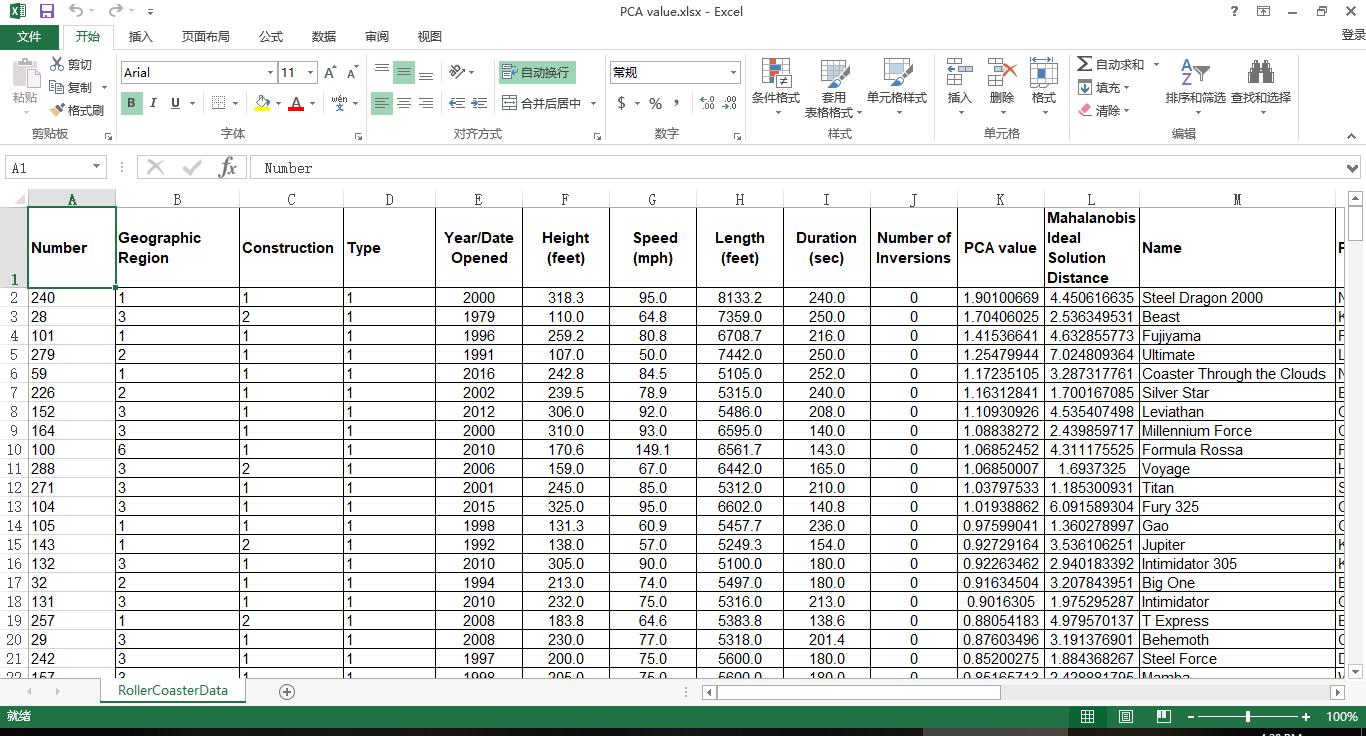
We also come up with a way that we can get the maximum value of year opened, height, speed, duration, length, and number of inversions in the data, setting them as an ideal solution. We then calculate the Mahalanobis distance between each data and the idea solution, taking the advantage of the avoidance of the effect of the dimension. The formula is shown as the following formula 2:



(2)

Among the formula, and denote two row vectors; denotes the covariance matrix; denotes the obtained Mahalanobis distance of the data. Part of the results is shown in the table 2, the rest of which are in the appendix.

Table 2: Ideal Solution result



However, we find that it is flawed for us to set the highest score as the ideal one, since no evidence guaranteed us that the statement is true. We still need to consider other methods.

* 1. **Principal Component Analysis**

With the help of PCA, we are able to rank the roller coasters.

We utilize the 9 original variables mentioned in 3.3 as the original data. We still use to denote independent variables matrixes and the dependent variables. The original variables are ; the new variables are . We use to denote the number of samples and use to denote the number of variables in each sample. Thus, the data matrix is as matrix 3[13]

(3)

Since the data vary in dimensions and ranges, we need to standardize the data. We adopt the variance standardization technique to operate the data so that the variance of the standardized data is 1, while we conduct the central translation so that the mean of the data is 0. The formula is as formula 4-5

denotes the standardized data at row and column ; denotes the data at row and column before standardization. denotes total column number and denotes total row number.

Then we establish the correlation coefficient matrix . The formulas are shown in formula 6-7.

(4-5)



(7)

(6)

Then we obtain the characteristic roots which satisfy for and characteristic vectors to determine the load on each new principal component variables of the original variables , which are equal to the larger characteristic values of the correlation matrix corresponding to the eigenvectors. is the value of the characteristic vectors. The formula is as formula 8:

(8)

In the formula, denotes each characteristic vector, denotes each characteristic value. The characteristic roots are shown in table 3. Characteristic vector matrix is in the appendix.

Table 3: Principal Component Analysis Characteristic Value

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 0.045393 | 0.232633 | 0.438638 | 0.607466 | 0.853558 |
| 1.011198 | 1.468301 | 1.707678 | 2.635135 |  |

The contribution rate formula and the total contribution rate formula is as formula 9-10.

and

(9-10)

We obtain the total contribution rate until the fifth principal component is 85.29%, which is larger than 85%. Therefore, we take the first fifth eigenvalue as the principal component. Suppose the principal component is formula set 11

(11)

In accordance with the first 5 scores of the principal component, we use the contribution rate as the weight and obtained the total score of each of the 293 roller coasters. Ranking the roller coasters, we put the first 5 roller coasters in table 4, the rest of which can be seen in the appendix.

Table 4: First 5 roller coasters of PCA analysis

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Number** | **Geographic Region** | **Construction** | **Type** | **Year/Date Opened** | **Height (feet)** |
| 240 | 1 | 1 | 1 | 2000 | 318.3 |
| 28 | 3 | 2 | 1 | 1979 | 110.0 |
| 101 | 1 | 1 | 1 | 1996 | 259.2 |
| 279 | 2 | 1 | 1 | 1991 | 107.0 |
| 59 | 1 | 1 | 1 | 2016 | 242.8 |
| **Number** | **Speed (mph)** | **Length (feet)** | **Duration (sec)** | **Number of Inversions** | **PCA value** |
| 240 | 95.0 | 8133.2 | 240.0 | 0 | 1.901007 |
| 28 | 64.8 | 7359.0 | 250.0 | 0 | 1.70406 |
| 101 | 80.8 | 6708.7 | 216.0 | 0 | 1.415366 |
| 279 | 50.0 | 7442.0 | 250.0 | 0 | 1.254799 |
| 59 | 84.5 | 5105.0 | 252.0 | 0 | 1.172351 |

Searching the top roller coasters online in the ranking of us [12], we find that all of the top 10 roller coasters online ranked the top one-third of us except the ones that cannot be found in the data set. Several top 10 coasters online are in the top 20 coasters of us. Thus it shows that the result online can be used as the training set. We download the scores from Costerbuzz [12] and use them for further modeling.

1. **Modeling**
   1. **Basic Statistics**

Figure 4: Line Chart of Duration. The Duration focus on 100-150 seconds interval.

After obtaining the original data, we do the basic statistics process. We download the score from the website, Coaster buzz, and set it as the dependent variables, while the variables given in the chart as independent variables. On the one hand, we make pie charts, as well as line charts, reveal the proportions of the roller coasters with each characteristic over the ensemble, as shown in figure 4-5.

Figure 5: Pie Chart of Geographic Region. The roller coaster from North America takes a major proportion.

The previous charts demonstrate, for instance: most of the given roller coasters locate in North America. The duration concentrates in 100-200 seconds interval.

* 1. **Analytical Hierarchy Process**

In order to choose by diverse factors and judge the condition of roller coasters, we utilize the Analytic Hierarchy Process (AHP) to achieve the goal which is to determine the weight of each option in complicated and uncertain problems. We define the 9 properties of the roller coasters as the scheme layer, while defining the score as the target layer, to build up the AHP model with three layers. The structure diagram 6 is shown as follows. [14]

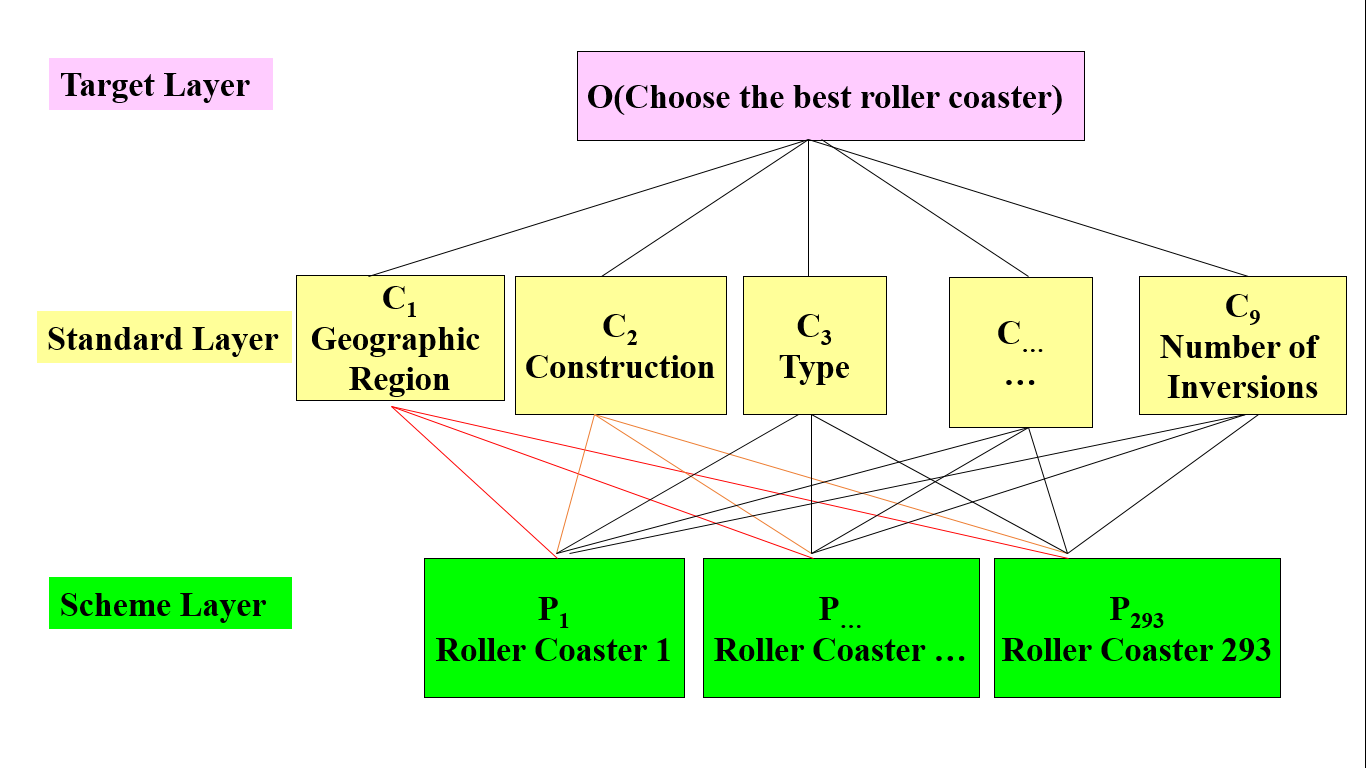


Figure 6: Structure diagram.

First, we define the amounts of roller coasters that possess certain properties under certain types of properties, which refers to the amount of a certain target choice under a certain scheme layer condition, as . In accordance with the target choice, we obtain a weight vector ( stands for the number of choices of target layer). We compute the ratio between the number, , of each scheme layer choice under a common target layer choice and regard it as the weight of paired comparison matrix. As they are consistent matrixes, we do not need to apply consistency tests to the matrixes, for they are automatically consistent, which means that the eigenvalues are all identical. With the help of the formula of the eigenvalue and eigenvectors shown in formula 12,

(12)

we can obtain the eigenvectors, . Composing the eigenvalues of each scheme layer, we obtain the eigenvector matrixes as well as weight vector matrixes from the target layer to the standard layer. The following tables 5-6 respectively shows the paired comparing matrix and eigenvector.

Table 5: Paired Comparing matrix from standard layer to object layer

|  |  |  |  |
| --- | --- | --- | --- |
| 1 | 2.866024 | 2.306462 | 1.349377 |
| 0.348915 | 1 | 0.80476 | 0.470819 |
| 0.433565 | 1.242606 | 1 | 0.585042 |
| 0.741082 | 2.12396 | 1.709278 | 1 |

Table 6: Paired Comparing matrix from standard layer to object layer

|  |
| --- |
| 0.396265 |
| 0.138263 |
| 0.171807 |
| 0.293665 |

Then we repeat the process from standard layer to scheme layer, and obtain a matrix of weight vector. Multiplying the two weight matrixes, we obtain the final weight matrix, which is the weight vector from scheme layer to target layer.

To define the paired comparison matrix from the standard layer to the object layer, we calculate the correlation coefficients between the online score and each given standard of data and the cross-ratio between the correlation coefficients. We discover that the possible value of Geographic Region varies too less, which means there are only two values that are different from the rest in the data with online score. The numbers of inversions exist too much zeros. The correlation coefficients of construction, type, and duration are too low for further analysis. Thus we merely take four standards to do further analysis, which are Year, Height, Speed, and Length, discarding the rest variables.

Finally, we draw the statistical chart with each weight vector, such as scatterplot, to clearly express the weight of the result of the roller coasters. The charts are shown in figure 7.

Figure 7: Result analysis. The weight of the result is irregular.

We can clearly see that the weight of each roller coaster is irregular, thus we need to consider a different method to the problem. We can infer that the irregularity may result from the low correlation coefficients.

* 1. **Linear Regression**

The third modeling method we use is Linear Regression. We can regard the properties of roller coasters as independent variables, and the online scores as dependent variables. Based on the samples, each data can be viewed as a mapping from the independent variables, which are the properties, to the dependent variables, which are sales. As each information is expressed numerical, we can find the function from the independent variables to the dependent variables through linear regression from the data. [15]

Let to respectively denote the nine properties respectively. Let denotes online scores. The value of the independent variables and dependent variables is the numbers of each option. We utilize regression formula 13.

(13)

Let denotes the independent variables matrix;denote dependent variables matrix; denotes coefficient matrixes. We apply Least Square Regression Method to the issue, of which the formula is shown in formula 14:

The formula is set to solve out the value of the coefficient matrixes of point estimation. With MATLAB giving solution, we obtain the coefficient matrixes which are presented in table 7:

(14)

Table 7: Linear Regression Coefficient

|  |  |
| --- | --- |
|  | -18.2625 |
|  | 0.089799 |
|  | 0.08964 |
|  | 0.017563 |
|  | 0.010858 |
|  | -0.00166 |
|  | 0.006959 |
|  | 8.52E-05 |
|  | -0.00053 |
|  | -0.01581 |

Point estimation possesses a drawback that it cannot express the accuracy of the data obtained. Thus we utilize interval estimation to reuse the Least Square Regression Method, the formula as in formula 15:

(15)

(16)

denotes the parameters to be estimated of the ensemble; denotes probability; denotes Confidence upper limit; denotes Confidence lower limit; denotes reliability which satisfies . In this way, we obtain formula 16

With the MATLAB program, we set as 0.95, under which the regression coefficient bound is shown in table 8.

The residual graph is shown in figure 8. When examining correlation coefficients, we find the correlation coefficients are 0.336705.

Table 8: Linear Regression Coefficient Bound

|  |  |  |
| --- | --- | --- |
|  | Lower Bound | Lower Bound |
|  | -28.9973 | -7.52761 |
|  | -0.35907 | 0.538668 |
|  | -0.07901 | 0.258286 |
|  | -0.04829 | 0.083415 |
|  | 0.005448 | 0.016268 |
|  | -0.00453 | 0.001203 |
|  | -0.00717 | 0.021093 |
|  | ######## | 0.000171 |
|  | -0.00273 | 0.00167 |
|  | -0.04853 | 0.01691 |

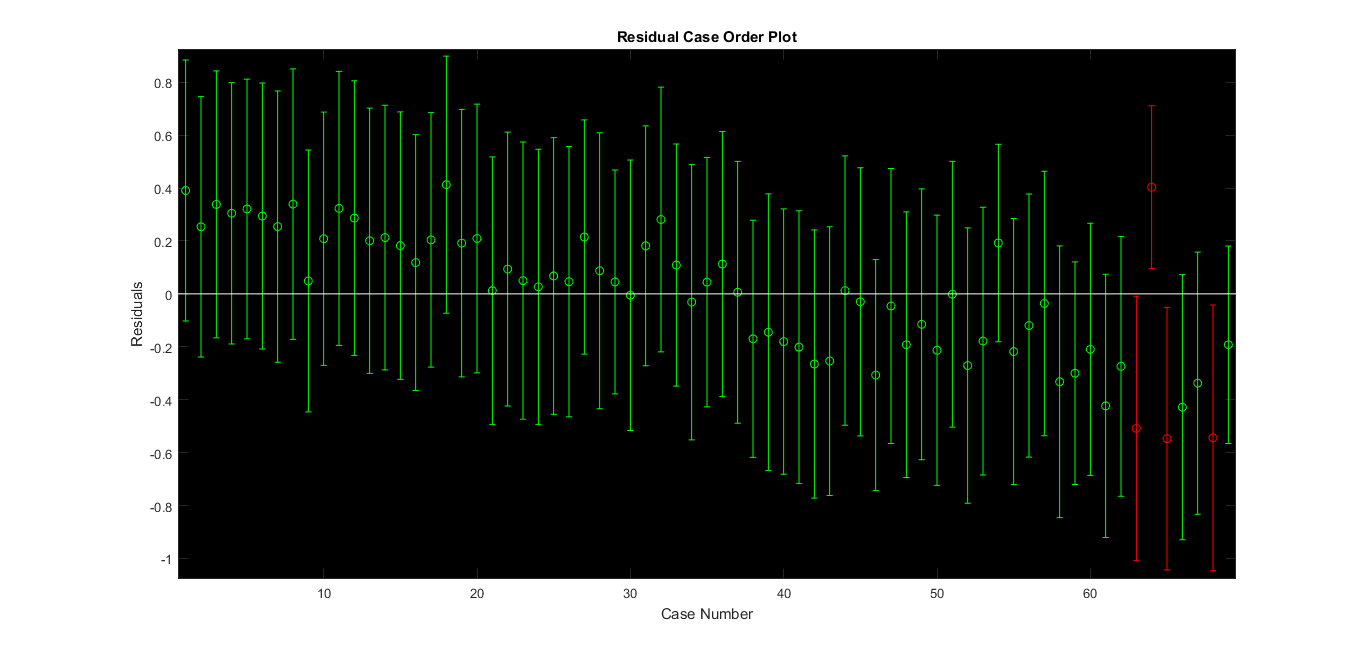


Figure 8: Residual Case Order Plot of Linear Regression

In light of the fact that the accuracy is relative low, which is insufficient to reveal the features of each variable precisely, we consider taking the advantage of other methods.

* 1. **KNN Algorithm**

In accordance with the given data, we try to use the data of which the online scores are matched to conduct the KNN algorithm to highly merge the vast amount of the data and find the shared features and characteristics of each sample to obtain the common properties of the roller coasters under similar condition to determine the relationship. [16]

We utilize Mahalanobis distance distinction to operate these data, which is processed after principal component analysis and features eradicating the dimension of each independent variables. The formula is as the following formula 17.



(17)

Among the formula, and denote two row vectors; denotes the covariance matrix; denotes the obtained Mahalanobis distance of the data.

For the accuracy, we correctly categorized 57 samples out of 69, achieving an accuracy of 83%. We made an optimization in 5.2.

1. **Optimization**
   1. **Principal Component Regression**

Principal Component Regression suits explicitly for the problems that have a vast amount of independent data types, not all of which are tightly connected to the dependent data, which means some of the data are loosely related to the data. In view of considering that our problem has 26 independent variables, the method is highly compatible with our research.

We can still do as part 4.3, regarding the properties of roller coasters as dependent variables and the online scores as independent variables. We try to reduce the dimensionality, diminishing the vast amount of the original data and variables into fewer data and variables, while the new variables can retain the information in the original data by and large. [17]

We utilize the 9 original variables mentioned in 3.3 as the original data. We still use to denote independent variables matrixes and the dependent variables. The original variables are ; the new variables are . We use to denote the number of samples and use to denote the number of variables in each sample.

Applying Least squares regression, point estimation and interval estimation method which has previously been mentioned, we obtain the principal coefficient matrix as shown in table 9 with formula 18.

(18)

Table 9: Coefficient Matrix of principal component

|  |  |  |
| --- | --- | --- |
| Point Estimation | Interval Estimation | |
| -17.9038 | -28.228 | -7.5796 |
| 0.017885 | 0.002086 | 0.033684 |
| -0.01057 | -0.01941 | -0.00174 |
| 0.013596 | -0.00226 | 0.029449 |
| 0.078525 | 0.004955 | 0.152094 |
| -0.02736 | -0.06604 | 0.01132 |

The correlation coefficients of this method are 0.322210105118927. Although there is no discernable elevation in the coefficient, the method focus more on the principal variables.

Ultimately, we conduct the inverse standardization process and obtain the equation interpreted in the original data, which is formula 19, and the final coefficient matrix, as shown in table 10.

(19)

Table 10: Final Coefficient Matrix of original variables

|  |  |  |
| --- | --- | --- |
| Point Estimation | Interval Estimation | |
| -17.9038 | -28.228 | -7.5796 |
| 0.075562 | 0.024277 | 0.126847 |
| 0.033707 | 0.016531 | 0.050884 |
| 0.020766 | 0.019247 | 0.022286 |
| 0.010701 | -0.05472 | 0.076122 |
| -0.00221 | 0.007277 | -0.0117 |
| 0.008843 | 0.01558 | 0.002106 |
| 0.000105 | -0.00306 | 0.003271 |
| -0.00103 | -0.01252 | 0.010466 |
| -0.00745 | -0.00913 | -0.00578 |

* 1. **Bayes Distinction**

Bayes Distinction ideally satisfies the requirements of such issue that each individual of the ensemble exists at different frequencies, which indicates that we need to take into consideration that the different possibilities that each individual exists. As for our research, each roller coaster is obviously impossible to appear at identical frequencies, so we apply Bayes Distinction to our study.

In the distance distinction method above, it does not take into account the frequency of each sample in the whole and does not take into account the loss caused by the wrong distinction. The Bayes distinction method modifies on the basis of distance distinction, and the formula is defined as in formula 20: [18]



(20)

Among which represents a posteriori probability; represents a prior probability; represents the frequency at which the sample appears; represents the total covariance matrixes. The distinction rule is that the posterior probability is the highest and the average wrong distinction loss is the lowest, which brings out the rule is as follows: If the condition meets the following formula 21:



(21)

Then we categorize into , among which is the ensemble, is the probability density function of , is prior probability of , which is the probability that it belongs a certain category when sample occurs, and is the number of . The solution formula for distinction analysis is as the following formulas 22-23:



(22)

(23)



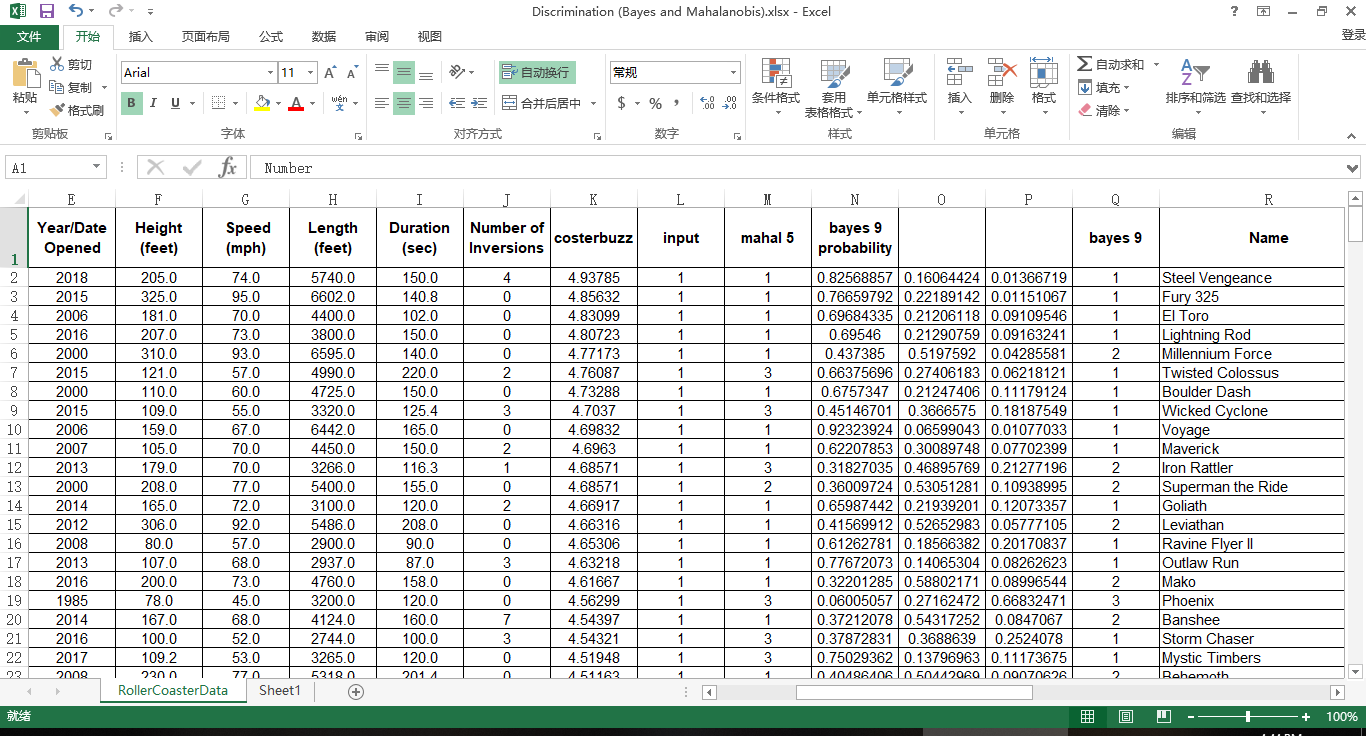
In this case, represents the conditional probability of wrongly categorizing the sample of to the ensemble . is the loss caused by this categorization. is a division of a set of distinction samples. is the average wrong distinction loss. The solution to a Bayes distinction analysis is to make the smallest set of solutions.

We divide the result of Bayes distinction into 5 categories, which are less than 4, 4 to 4.5, and 4.5 to 5. For the training set, if the online score lies in 4.5 to 5, we define the roller coaster as category 1. Likewise, we define the roller coaster of which the score is from 4 to 4.5 as category 2. We randomly pick out a certain amount of data from ALL the data which has no score online or the score is lower than 4 and define them as category 3. Using the MATLAB program, we still use all the data with online score to carry out Bayes distinction solution.

The result is shown in the appendix, part of which is as following figure 8-9 and table 20. For instance, the number “36” shows that there are 36 samples with sit down type are judged as Category 1, which is the high score category.

For the accuracy, we correctly categorized 59 samples out of 69, achieving an accuracy of 85%, which is relatively higher than the accuracy obtained from KNN algorithm. The following table 11 is a part of the result.

Table 11: Bayes Distinction Result



We also make various charts and tables to exhibit our results, part of which are as the following figures 9-10 and table 12:

Figure 9: Bayes Result of Construction in Low Score

Figure 10: Bivariate plot from Height to Bayes Category

Table 12: Bayes Category to Type

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Sit down | Inverted | Stand up | Suspended | Flying | Wing |
| Category1 | 36 | 0 | 0 | 0 | 0 | 0 |
| Category2 | 59 | 17 | 2 | 2 | 5 | 3 |
| Category3 | 132 | 21 | 0 | 4 | 0 | 1 |

From the results given, we can clearly figure out the trend that the roller coasters which are in the place far away from North America tend to have a high score, especially the ones locate in Middle East, Oceana, and Russia. The roller coasters that are made from wood are more likely to have a higher score. A newly opened roller coasters are more welcomed. If the roller coaster is relatively higher, it is more possible to achieve a better score. 2 and a half minutes and 60 mph are a proper time for a loop and a satisfactory speed respectively. If the number of inversions is too high, it may conversely do harm to the passion of tourists to ride.

* 1. **BP Neural Network Fitting**

BP Neural Network is a kind of multilayer feed-forward network, which highly fits for the problem that there are data with a certain scale, the relationship between which is not too complicated to identify. When it comes to our target, we have a middle-sized database, while the process we want is fitting, which is not too intricate, which shows that the model can be applied to our goal.

We utilize BP neural network fitting as another method to promote the accuracy of the regression. BP neural network works to encode itself with its high-dimensional features and to carry out dimension reduction processing towards high-dimensional data. It is marked by a feature extraction model with unsupervised learning, which can also combine a few basic features to obtain higher-layer abstract features. [19]

We utilize Tangent Sigmoid function as the transfer function; we use Levenberg Marquardt algorithm (trainlm) as the training algorithm; we use the Gradient descent with momentum weight and bias learning function (learngdm) as the learning algorithm; we use the mean square error (MSE) method as the learning function. The structure of the network and the performance plot are shown in figure 11 and 12.



Figure 11: BP Neural Network Structure. The layer number, which is 20, does not consumes too much time while the result is satisfactory.

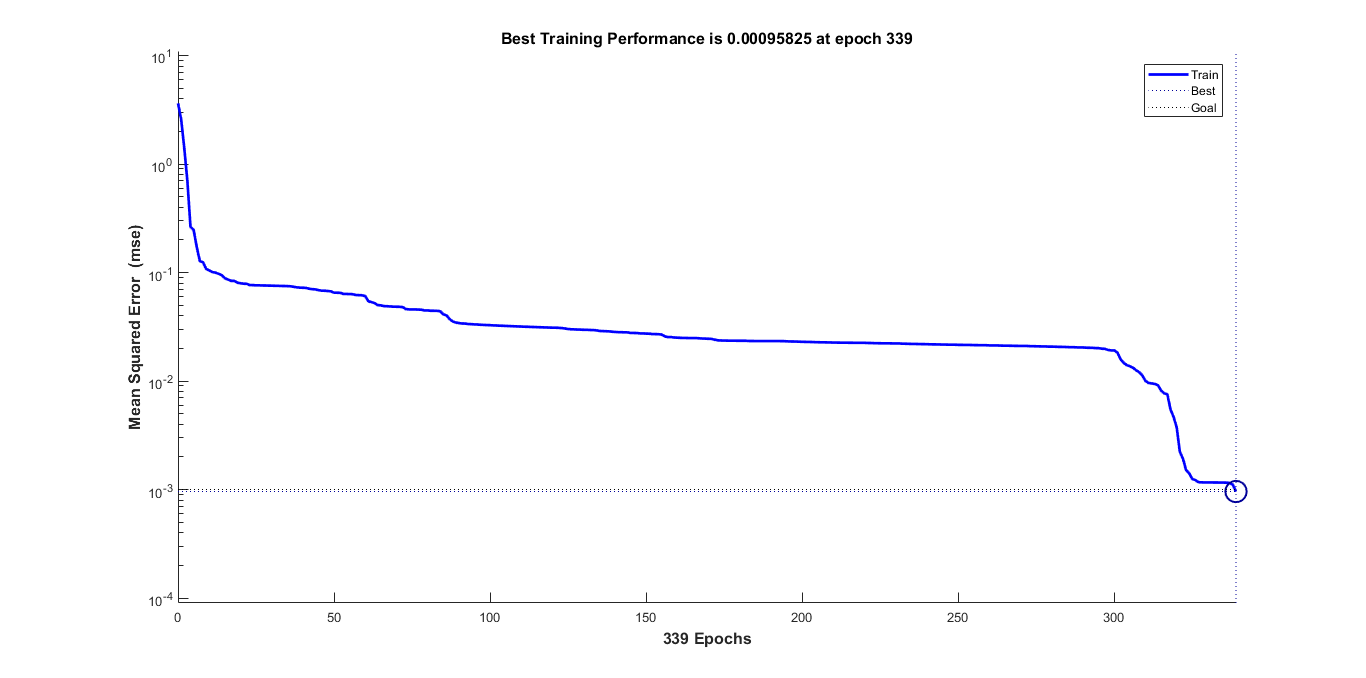
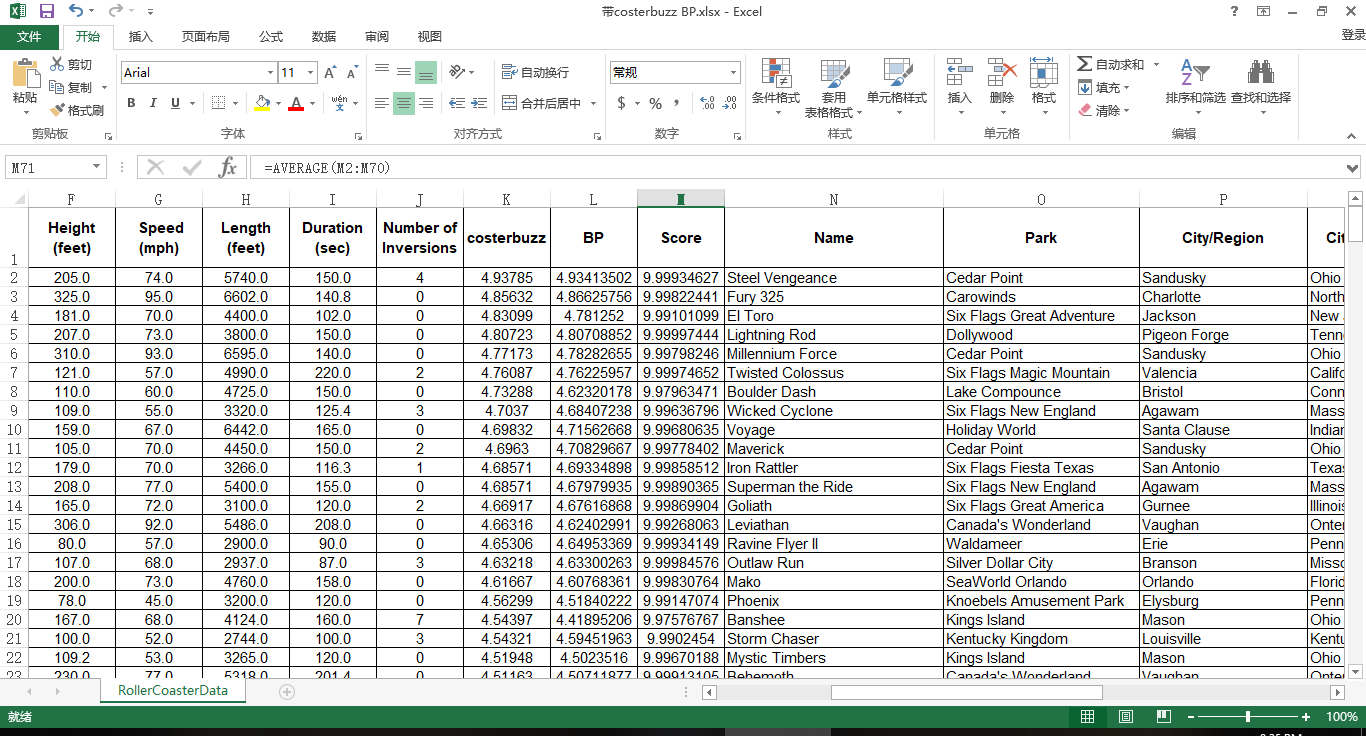


Figure 12: the performance plot of BP Neural Network. The training performance is enhancing rapidly.

Using the MATLAB program, we still use all the data with online score to carry out the BP neural network fitting.

We consider dividing the learning samples into three groups, each time using two of the groups to carry out a model and then test the test set. In light of the fact that there are a mere 69 training data, it is not sufficient enough for us to conduct in this way. Hence, we use all the training data to training the BP Neural Network Algorithm. The result is in the appendix, part of which is as the following table 13.

Table 13: BP Neural Network Result. The error of some numbers is lower than 1%.



It can be seen that some of the predicted data run an accuracy that is higher than 99%.

* 1. **XG Boosting Algorithm**

We utilize XG Boosting algorithm to obtain the average value of each method of the samples. The basic formula is as the following formula 24

(24)

In the formula, denotes the overall loss of each classifier, denotes each classification function, and is a function of each classifier to reduce the loss. denotes the original result of Principal Component Analysis. denotes the result of Bayes distinction. denotes the original result of BP neural network fitting. For each category in Bayes distinction, we utilize the mid-value of each interval to numerate each category. We divide the result of Bayes distinction into 5 categories, which are less than 4, 4 to 4.5, and 4.5 to 5. Therefore, we use 3.75, 4.25, and 4.75 to denote the 3 result of the categories.

The main theory of BOOST algorithm is as follows. For a complicated issue, it is a better judgment when synthesizing the judgment of each expert than that of a sole expert. For each step, we generate a model accumulate each model to a whole model, which enables us to analyze the problems. Hence, we need to assemble several weak learner into a strong learner by determining the loss functions, , to minimalize the error and loss of misjudgment.

We input the predicted result of the three learner into the algorithm as the learning set and the real result as the target goal. We regard test set in the Bayes distinction and BP Neural Network as the testing set. With the help of XG Boosting module in PYTHON, we are able to determine the weight of the three learner to generate the final result. [20]

We utilize a formula to measure the error of our estimation, reaping an average score

same as the original result, receiving almost a full score of 10, which shows that this model can successfully reflect the trend. The formula is as the following formula 25.

(25)

In the formula, denotes the score of the data, while and respectively denote the predicted value and the real value of the data.

We discover that many roller coasters have the same value of XG Boosting, which may due to the reason that there are too less training set while too much testing set. Since the BP Neural Network reap a extremely accurate outcome, we decide to use the result of the BP Neural Network as the final score and ranking by BP Neural Network if the outputs of XG Boosting are identical.

1. **Sensitivity Analysis**

Sensitivity analysis is a method of studying and analyzing the sensitivity of the model to changes in system parameters or surrounding conditions. In the optimization methods of our team, it can detect the stability of our model, especially when the given data is not accurate.

In this part, we will mainly discuss the sensitivity of the application part. If we give the test set of the data an increase or a decrease of 1%, by changing the value of the original data matrix on the program, we discover that the output data of the principal component regression changes precisely 1%; almost all the results in the Bayes Distinction part have no difference in categories; the majority of the output of BP neural network model fluctuates 1% approximately. The output after the change is small enough for us to make a further adjustment. Therefore, it is acceptable in the modeling. This sensitivity analysis also indicates that our model has universality and can be applied to more situations. For instance, if there is some error in the data, out final result does not vary rapidly correspondingly. Therefore, our model is relatively stable. The data of Sensitivity Analysis can be referred to the appendix part.

1. **Conclusion**
   1. **Strength and Weakness**

The strength of the model for the rating of roller coasters and the algorithm being used mainly include the following aspects:

The methods applied in this model includes both qualitative and quantitative analysis, and different conclusions from various methods can be obtained through our modeling process. As for the qualitative analysis, for instance, from the results of Bayes Distinction, we can figure out the trend that the roller coasters which are in the place far away from North America tend to have a high score and that the roller coasters made from wood are more likely to have a higher score. A newly opened roller coasters are more welcomed. If the roller coaster is relatively higher, it is more possible to achieve a better score. 2 and a half minutes and 60 mph are a proper time for a loop and a satisfactory speed respectively etc. All of these are obtained through the qualitative analysis which provide us with valuable information. As for the quantitative analysis, it is obvious that nearly all the algorithms being applied need the input of the data base and the whole rating process our model depends on needs the analysis and testing of quantitative property of every roller coasters. Far from the common evaluating methods based on the riders' subjective inputs currently, our model is based on facts, analysis and trainings, which will definitely provide more accurate and scientific results on the ranking.

Our methods take into consideration both continuous and discrete variables, the fact of which make our model especially suitable for the question's requirement. For instance, in our optimized model, the data of discrete independent variables, like the material used for construction of the roller coasters, are mainly analyzed by Principal Component Analysis, which is suitable for the processing of discrete data. On the other hand, Bayes Distinction mainly served to analyze the continuous variables including average speed, maximum height, etc. For the XG Boosting algorithm, it is suitable for all independent variables and can successfully synthesize the result of the optimized model. Thus, our model has sufficiently taken into account the property of different types of data, and the result yielded will in turn be highly persuasive.

Another outstanding point of the whole modeling process is the variety of methods being used---from the basic statistics and linear regression to the optimized model of BP Network Fitting and XG Boosting algorithm. The final ranking is produced through exploration on various methods, and we have been continuously evaluating the viability of different models and thinking about the how to further improve our results through more advanced methods. It is our strive for excel that guarantees a more accurate and suitable method as a whole for the quantitative analysis of the rating. The multiplicity of methods not only shows our clear logic chain from the perspective of pragmatic problem solving, but also ensures a more stable and precise result.

The model is also propagable as an common method for the rating of any new roller coaster given. In other words, beside the 300 roller coasters provided in the table, if any new roller coaster with basic properties given are added to the data base of our whole modeling process, a rating can also be produced and be added to the original ranking.

Furthermore, our model is highly flexible and can be applied to various situation. For instance, if another new property, no matter continuous or discrete, is added to all the roller coasters, the method and logic behind of our modeling can also be applied, since the ranking is based on a whole series of analysis instead of randomly assigning weight to each property. Thus, a different ranking is likely to be produced. Besides, it is also worth noticing that if more roller coasters with ratings are added to the training set in the XG Boosting algorithm, a more accurate result will be produced. This is due to the fact that certain optimized methods like XG Boosting algorithm yield results based on the self-learning of data input, which indicates more data analysis and more source for learning will boost the accuracy and stability of the method. Therefore, the model we proposed can learn constantly from the additional data input, so it is highly flexible and makes possible the dynamic adjustments, which makes the model suitable for being applied under various circumstances.

Finally, our model yields the precise rating results instead of just the ranking of the top 10. Actually, it is also possible to produce the quantitative result of any roller coaster given. This will definitely be better than the vague ranking result which is a much weaker conclusion compared. The accurate rating can reflect the difference in a quantitative way between each roller coasters and give the potential riders more appreciable information.

The weakness of our model mainly includes the aspects following:

The data of the roller coasters given is sometimes not ample and sufficient enough for the evaluation of one particular independent variable. This will in turn cause the data input being inaccurate since some information is missing and cannot be applied to analysis. Though interpolation is done during the data cleaning process, there are still some data left vacancy because of the insufficiency of existed data, for the loss of large quantity of data will make the interpolation process meaningless. So the loss of data will affect our final rating, even though the methods we applied have considered to maintain the original information as much as possible.

Besides, there are only 300 pieces of information given for the learning and testing of data, the fact of which will unavoidably makes the method like XG Boosting not accurate enough. For methods like this, more pieces of data being learned will further boost the accuracy. However, despite some deficiency, it is still a suitable method for the synthesis of the optimized result and provide us with the relative more stable and precise rating result

.

* 1. **Comparison of the top 10 Roller Coasters**

To give out the final ranking of our model, we decide to use the result of XG Boosinig as the ranking criterion. As mentioned above, we find a flaw that many scores of roller coaster of XG Boosting are identical. Hence, it is difficult for us to distinguish which roller coasters should be the top ones. In light of the high accuracy of the BP Neural Network given in 5.3, we decide to use the result of the BP Neural Network to rank the roller coasters if the results of XG Boosting are exactly the same. In other words, the first ranking criterion is the result of XG Boosting, and the second one is the result of BP Neural Network. The top 10 roller coasters and their scores are shown as following table 14.

Table 14: Final Ranking

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Number** | **Name** | **Park** | **City/Region** | **City/State/Region** | **Country/**  **Region** | **Geographic Region** |
| 257 | T Express | Everland | Yongin-si | Gyeonggi-do | South Korea | Asia |
| 9 | Anaconda | Walygator Parc | Maizieres-les-Metz | Lorraine | France | Europe |
| 66 | Crazy Coaster | Loca Joy Holiday Theme Park | Yongchuan | Chongqing | China | Asia |
| 10 | Apocalypse | Six Flags America | Upper Marlboro | Maryland | United States | North America |
| 33 | Big Thunder Mountain | Disneyland Resort Paris | Marne la Vallee | Ile-de-France | France | Europe |
| 273 | Tonnerre de Zeus | Parc Asterix | Plailly | Picardie | France | Europe |
| 143 | Jupiter | Kijima Kogen | Beppu | Oita | Japan | Asia |
| 59 | Coaster Through the Clouds | Nanchang Wanda Theme Park | Xinjian | Nanchang, Jiangxi | China | Asia |
| 87 | Firehawk | Kings Island | Kings Mills | Ohio | United States | North America |
| **Number** | **Inversions (YES or NO)** | **Status** | **Construction** | **Type** | **Drop (feet)** | **Year/Date Opened** |
| 257 | NO | Operating | Wood | Sit Down | 150.9 | 2008 |
| 9 | NO | Operating | Wood | Sit Down | 40.0 | 1989 |
| 66 | YES | Operating | Steel | Sit Down |  | 2013 |
| 10 | YES | Operating | Steel | Stand Up | 90.0 | 2012 |
| 33 | NO | Operating | Steel | Sit Down | 39.3 | 1992 |
| 273 | NO | Operating | Wood | Sit Down |  | 1997 |
| 143 | NO | Operating | Wood | Sit Down |  | 1992 |
| 59 | NO | Operating | Steel | Sit Down | 255.9 | 2016 |
| 87 | YES | Operating | Steel | Flying |  | 2007 |
| **Number** | **Height (feet)** | **Speed (mph)** | **Length (feet)** | **Duration (min:sec)** | **Duration (sec)** | **Number of Inversions** |
| 257 | 183.8 | 64.6 | 5383.8 |  | 138.6 | 0 |
| 9 | 118.1 | 55.9 | 3937.0 | 2:10 | 130.0 | 0 |
| 66 | 108.3 | 52.8 | 2870.8 |  | 178.4 | 10 |
| 10 | 100.0 | 55.0 | 2900.0 | 2:00 | 120.0 | 2 |
| 33 | 72.2 | 40.4 | 4921.3 | 3:56 | 236.0 | 0 |
| 273 | 98.0 | 52.0 | 4044.0 | 2:05 | 125.0 | 0 |
| 143 | 138.0 | 57.0 | 5249.3 | 2:34 | 154.0 | 0 |
| 59 | 242.8 | 84.5 | 5105.0 | 4:12 | 252.0 | 0 |
| 87 | 115.0 | 50.0 | 3340.0 | 2:10 | 130.0 | 5 |
| **Number** | **BP** | **XGBoosting** | **G Force** | **Vertical Angle (degrees)** |  |  |
| 257 | 6.706793 | 4.68571 |  | 77 |  |  |
| 9 | 6.004326 | 4.68571 |  |  |  |  |
| 66 | 5.979574 | 4.68571 |  |  |  |  |
| 10 | 5.926418 | 4.68571 |  |  |  |  |
| 33 | 5.87154 | 4.68571 |  |  |  |  |
| 273 | 5.777834 | 4.68571 |  |  |  |  |
| 143 | 5.754164 | 4.68571 |  | 45 |  |  |
| 59 | 5.726858 | 4.68571 |  |  |  |  |
| 87 | 5.54524 | 4.68571 | 4.3 |  |  |  |

Apart from the score online we use in the modeling part, we find a second scoring website from the website Coaster Critic [21], of which the score can be referred to the appendix. Comparing the two top 10 roller coasters, the biggest difference we find is that out scores focus on not only the roller coasters in the US, but also the roller coasters in a world scale. The location of our top 10 roller coasters include the US, France, Korea, China, and Japan, which demonstrates that we truly achieve the goal that selecting the roller coasters based on objective and quantitative data rather than personal, subjective opinion. We are able to recommend the roller coasters only from the properties of themselves rather than personal opinions.

There are also some consistency between the two scores. For instance, wood roller coasters are both highly rated, of which the reason may lies at people prefer the obsolescence of conventional wood roller coasters. The top 10 roller coasters of both are less likely to have inversions comparing with the roller coasters ranked after 10. The opening years both cover a wide rage, from the 1980s to two years ago. The roller coasters with middle speed, length, duration, or height are both leading the top of the rank.

We also compare our result with other website, such as the result from MostLuxuriousList or TheTopTens®, of which the result can be seen in the appendix (some roller coasters in the two websites are missing in the given database of roller coasters). The case are also similar to the previous website. The top roller coasters online concentrate in the US, while the top coasters our model gives out involves a wider range. It is mutually revealed that the parameters of the top coasters lie in middle interval, which manifests that middle-interval coasters are more warmly welcomed.

* 1. **Concept and design for a user-friendly app**

The user-friendly app we construct mainly aims to satisfy the riders' needs on roller coaster riding selection and meet individual demands. The app mainly contains 3 aspects--- recommendation of roller coasters based on all the applied riders' experience on a global scale, the specific recommendation of roller coasters to individuals after the data processing and the analysis on the individual's past preference, and the selection of roller coasters by the filter to meet the users' needs. The roller coasters' own prosperities form the basic data base for the selection and recommendation, and the algorithm will help with the analyzing process.

To begin with, the app will ask for the individuals' personal information including the region they live in. Then it can first select the roller coasters from that region and make corresponding recommends. The registered riders will be required to rate the roller coasters they have rid after the thrilling experience, and each piece of information they record will be put in the data base for the analysis. To encourage the registered riders to make contribution to the data base, some rewards maybe provided. The question may involve the following aspects: the feeling after the ride, the degree of excitement and stimulation based on the individuals' experience, the rating of roller coaster as a whole, etc. All of these questions are the users subjective inputs, based on the rating they have given, we can use them as the training set in the XG Boosting algorithm mentioned above and use them to refresh the ranking of the roller coasters at every moment. In this way, the subjective information can be turned as input of quantitative analysis, improving the original model's accuracy and stability. At the same time, this ranking would be used to provide the new users with the top roller coasters and encourage them to experience the best ride.

The information provided constantly by one individual---the track record---can also provide useful information on the individual's own preference. The basic algorithm behind this function is try to determine the correlation of the riders' record and the roller coasters in the data base. To define the similarity between the historical data of the users and the roller coasters, we can set each data of the user or the roller coaster which the user has not rid yet as a row vector and calculate the correlation coefficients between the two. Then we can rank the roller coasters by the correlation coefficients from the largest one to the smallest one, recommending the one with the several largest correlation coefficients. We can also calculate the Mahalanobis distance between the row vectors previously mentioned and ranking as above, taking the advantage that the method does not take the dimension of the data into account. Based on the algorithm, the app can thus successfully achieve its second crucial function, and make the recommendation based on the quantitative analysis, spotting the users' need and saving the users' time for searching.

The app could also set up a selecting system to meet the riders' special needs. The system will be much like a search engine but it will be entirely based on the property of the roller coasters. To make the sifting process more user-friendly, the options for the potential riders to choose will not include specific numbers. For instance, if the potential rider want to select a roller coaster with longer duration time, the search engine will not require them to put in specific numbers, but only choose from different levels such as short(30-60sec), medium(60-120sec),and long(>120sec). Different selecting options will thus minimize the number of roller coasters based on the rider's demand and correspondingly make the proper recommendation.

Besides from the main purposes, auxiliary functions may also be included. First, a community will be set up to let the riders share their own riding experience, which may boost their sense of belonging with others who also like roller-coaster riding. They may even find the app useful as it can allow them to make friends with those who share the same interest with them. Besides, basic information of the roller coaster sites around the global will be provided, in the form of both pictures and videos to give the potential riders a real sense of spectacularity, and every rider is welcomed to write their own experience and comments. For those especially love the thrilling feeling, they can also keep a journal in this app, and write down anything they want to recall about every one of their stimulating experience. Furthermore, up-to-date news about the roller coasters around the world will be timely reported by converging the information online, capturing the riders interest and promote them to have a try. Some related commercial products like key chains and post cards could also be provided after the cooperation with certain entertainment companies.

In brief, the app we construct will use the algorithm and quantitative analysis to meet the potential riders' needs and help them decide the best option, guaranteeing them a satisfying and enjoyable experience.

1. **Reference**

[11] https://wenku.baidu.com/view/c6e8f183b9d528ea81c779af.html, Clustering Analysis with Matlab.

[12] https://coasterbuzz.com/RollerCoasters/Top100, Top 100 Roller Coasters: The CoasterBuzz 100.

[13] <http://blog.csdn.net/MATLAB_matlab/article/details/59483185?locationNum=10&fps=1>, MATLAB principal component analysis.

[14] Mingbei C., Gang H., Guoufu Z. Comprehensive evaluation of takeaway website based on AHP method——Eleme website as an example [J]. Modern Business, 2015, 12: 57-58.

[15] <https://wenku.baidu.com/view/99c8408e6529647d272852cd.html>, Interval estimation and linear regression analysis with MATLAB.

[16] Jiang W. Comparative Study of Fisher Discriminant and Mahalanobis Distance Discriminant [J]. Journal of Ningbo Polytechnic, 2017, 21(5); 91-94.

[17] Yimeng F. Analysis of influencing factors of customer purchase behavior in E-commerce [J], Industrial & Science Tribune, 2014, 13(8): 138-139.

[18] Haiwei W. Yu X., Yalin W.A bivariate hierarchical Bayesian approach to predicting customer purchase behavior [J], Journal of Harbin Engineering University, 2007, 28(8): 949-954.

[19] Wu P. Application of Cigarette Sales Forecasting Based on Neural Network [J], Computer Simulation, 2012, 29(3): 227-230.

[20] https://blog.csdn.net/weixin\_42029738/article/details/81675234, Hand-in writing XG Boost programs.

[21] http://www.coastercritic.com/roller-coaster-reviews/, Coaster Reviews List – CoasterCritic.

[22] https://www.mostluxuriouslist.com/top-10-best-roller-coasters-in-the-world/, Best Roller Coasters in the World - List of Top Ten Ranking.

[23] https://www.thetoptens.com/best-roller-coasters/, Top Ten Best Roller Coasters - TheTopTens®.

1. **Appendix**

Bllabla