1. **Dataset Introduction**

The data set used in this operation is the shared bike data set. Data sets from Kaggle https://www.kaggle.com/c/bike-sharing-demand. There are 10,886 rows of data in the dataset, which includes hourly weather information for each day, outdoor and felt temperatures, seasons and holidays and working days, and the number of non-registered users and registered users for each time period. Weather 1,2,3,4 means the weather is getting worse.

1. **Research Question Significance**

I mainly use this data set to provide some help to bike-sharing companies, so that they can make more reasonable decisions. I hope to make a model, sharing bicycle companies can predict the number of shared bicycles rental during this period through a large amount of data every hour, so that company can more rationally arrange the plan of car release, car maintenance and so on. To better serve customers.

In addition, cluster technology is used to analyse user behaviour patterns and understand the differences between registered and non-registered users, so as to develop better plans to attract new users.

1. **Machine Learning Technique**
2. Decision tree based on season, weather, working day, holiday and time data.
3. Multi-feature polynomial regression based on temperature, perceived temperature, humidity and time data.
4. K-Means clustering algorithm based on the number of registrants, non-registrants and total registrants.
5. **Machine Learning Models**

Decision tree

1. Observe the influence of discrete data on the target through the form of data visualization. I set the X-axis as a 24-hour period to show the impact of these discrete data on the number of bicycle rentals at each period of the day.

2.Then we build the decision tree model, calculate the mean square error and R-squared, and carry out visualization processing. The graph on the right shows our Mean Squared Error and R-squared, and we can see that the fitting effect of this model is quite good.

3. Then I take the average of these data points to generate a line chart. The actual data lines and the predicted lines are almost identical, suggesting that the model is more inclined to use time periods of the day for splitting, while other features (such as seasons, weather, working days, holidays) are less important.

4. I then used the parameter grid to select better parameters for the decision tree, and we can see that these parameters are more overfitting and less accurate for the test set prediction.

Multi-feature polynomial regression

Firstly, we analysed the rest of the continuous data temp, atemp, humidity and hour with linear regression, Drawing the mean line of the prediction, we can see that the prediction did not work very well.

In order to improve the model, we use multinomial multifeature linear regression of order 2, 5, and 10. We can see that the model has significant changes after the addition of polynomials. However, as the model continues to rise, the model will be too complex, although the performance on the training set is better, but due to overfitting, the effect on the test set is very poor. So, we compare and choose 5 as our degree.

K-Means clustering algorithm.

Finally, we use K-Means clustering algorithm to cluster the data set and analyse the non-registered user behaviour. Divided into two clusters. Users in cluster 1 are more likely to be non-registered users who prefer to use their bikes in relatively high temperatures, low humidity, and high wind speeds. The Silhouette Coefficient is 0.178. The clustering effect is normal, and the in-cluster similarity is similar to the inter-cluster similarity.

**6. Data Analysis and Results**

a. Decision tree can well predict the number of bicycle rental at every moment through season, weather, working day, holiday and time data.

1. The second model, polynomial regression, can also predict the number of bike rentals through temperature, perceived temperature, humidity and time data.
2. Finally, K-Means clustering can understand the behaviour patterns of non-registered user groups, This may indicate that non-registered users are more inclined to use bicycles in hotter weather and relatively crisp conditions.

**7. Conclusion**

Decision tree regression for discrete data can capture nonlinear relationships and can better adapt to complex patterns. I have also mastered the use of parameter grid, so I can better find the best parameters. In terms of polynomial regression, we know the advantages of polynomial regression and the choice of polynomial order. Master K-Means clustering algorithm, cluster different behaviour patterns in unsupervised learning, and deeply understand the user behaviour patterns of different clusters.

**8. Limitations Discussion**

Decision tree is easy to overfit and be easily disturbed by noise. Moreover, combining discrete data with linear data to make a model, random forest is a good choice.

There is also a K-means cluster analysis that finds the best K value without any more swapping the number of clusters. Several cluster analysis methods should also be used to compare their advantages and disadvantages.

supplementary instruction

1.Dataset Introduction

## Data Fields

**datetime** - hourly date + timestamp    
**season** -  1 = spring, 2 = summer, 3 = fall, 4 = winter   
**holiday** - whether the day is considered a holiday  
**workingday** - whether the day is neither a weekend nor holiday  
**weather** - 1: Clear, Few clouds, Partly cloudy, Partly cloudy  
2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist  
3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds  
4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog   
**temp** - temperature in Celsius  
**atemp** - "feels like" temperature in Celsius  
**humidity** - relative humidity  
**windspeed** - wind speed  
**casual** - number of non-registered user rentals initiated  
**registered** - number of registered user rentals initiated  
**count** - number of total rentals

Data preprocessing:

When Weather=4, there is only one piece of data, which is too small, so it is deleted.

Converts the datetime column to the appropriate format.

4.Machine Learning Models

Decision tree

1. Observe the influence of discrete data on the target through the form of data visualization. I set the X-axis as a 24-hour period to show the impact of these discrete data on the number of bicycle rentals at each period of the day. This is the weather; you can clearly see the decline in the number of cyclists as the weather gets worse. The second is about the seasons, and we can see that there is a significant decline in the spring.

2.The third is the line chart of working days, we can obviously feel that the distribution of working days is mainly in the time of going to work and going home. What is markedly different is the holiday days, when we can see that there is no clear peak in the number of bike rentals during the holidays. This part of the data shows that the weather, season, working days, holidays and the number of rentals per hour have a clear relationship.

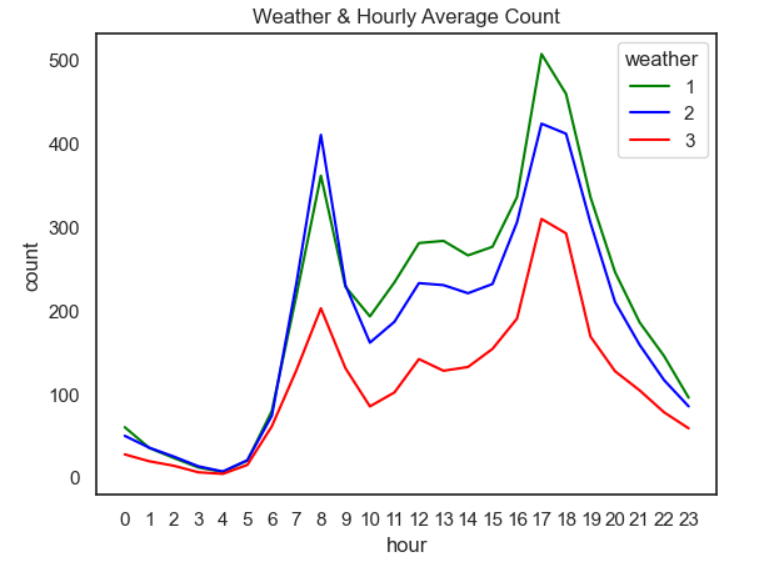
3.The following is a picture of the relationship between the number of people and each time period of each day within 24 hours. We can clearly see that there is a very strong correlation between the time and the number of people.

4.Then we build the decision tree model, calculate the mean square error and R-squared, and carry out visualization processing. The graph on the right shows our Mean Squared Error and R-squared, and we can see that the fitting effect of this model is quite good.

5.In this picture，The X-axis is the time period, the blue is the real data point of all the test data during this time period, and the pink is the predicted data point, with little difference.

6. Then I take the average of these data points to generate a line chart. The actual data lines and the predicted lines are almost identical, suggesting that the model is more inclined to use time periods of the day for splitting, while other features (such as seasons, weather, working days, holidays) are less important.

7.I then used the parameter grid to select better parameters for the decision tree, and we can see that these parameters are more overfitting and less accurate for the test set prediction.



表格

描述已自动生成

图形用户界面, 应用程序, Word

描述已自动生成data size

图形用户界面

低可信度描述已自动生成 表格

中度可信度描述已自动生成data completed

图形用户界面, 文本, 应用程序

描述已自动生成

data isn’t duplicated

图形用户界面, 文本

描述已自动生成 too small, so it is deleted.

图表, 折线图, 直方图

描述已自动生成图表, 折线图

描述已自动生成

图表, 折线图

描述已自动生成

图表, 折线图

描述已自动生成

图表, 直方图

描述已自动生成文本

描述已自动生成

图表, 散点图

描述已自动生成图表, 折线图

描述已自动生成Word

低可信度描述已自动生成

图表

描述已自动生成文本

描述已自动生成文本

描述已自动生成图表

描述已自动生成图表, 散点图

描述已自动生成图形用户界面, 文本, 应用程序

描述已自动生成