Lingge Wu:957116980
Junfei Sun:931760384
Jihui Yang:973811076

Machine Learning Group Project Report

Project Description

Overview:

In this project, we apply the concepts we have learned about machine learning to real-world problems. We use the related concepts of linear regression to predict the real-world problem - the number of shared bicycle rentals. And on this basis, a deeper analysis was carried out.

Overall goal:

The overall goal of our project is to predict the number of rental bicycles per day and also the input can be a period of time (like several days or one week)and will make a prediction to output the best day of the period to hang out for fun.

Implementation function overview

We will implement the following functions in this project:

- Processing and optimization of raw data
- Application to Regression Models
- Training and testing on the dataset
- Prediction

Algorithms may be needed:

This problem we currently think it as a Linear regression problem, and we will use the least squares method to build our training model and achieve our goal.

Our data set:

We found the data set in a public website. Here is the Link:

https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset

There are two date sets we may use to train or test: hour.csv and day.csv

Day.csv: totally 732*16 which includes 731 instances and for each instance there are 16 attributes.

Hour.csv: This dataset is more like to separate the day.csc by hour, and should be 732*24, but I found that there were some lines that were missing.

Some simple ideas for designing:

We first will start to process the data set and avoid the attribute which is not linear related to make sure it can satisfy the requirements to build the Linear regression model. And then split the dataset and distribute 80% to training dataset and 20% to the testing dataset.

Use least squares method to build the model(We final use the python package sklearn.linear_model import LinearRegression,since there is a unfound error with least squares)

Calculate the accuracy

Draw diagram

Make an prediction the best day hang out

Python package may needed:

Matplotlib: draw diagram

Numpy: use to calculate matrix

Sklearn: build model

Pandas: load the dataset

Programming tools:

Anaconda: A collection of libraries for easy installation of various runtime libraries

Pycharm: Integrated Development Environment for python

Github: A code hosting platform, which is conducive to team members participating in code writing

Group member:

Lingge Wu: My role on the team is to work on the dataset. There are many factors in the dataset to ensure that the training results are as accurate as possible. I want to throw out the less correlated factors. Here I use two ways. The first way is to judge the linear relationship between the independent variable Xi and the target value by drawing a scatter plot. If the linear relationship shown in the graph is poor, I would consider removing its Xi from the dataset. The second method is to find the correlation coefficient between all factors and the target value. The correlation coefficient is a value between 0 and 1 in absolute value. The closer the value is to 1, the higher the correlation. Here I consider removing elements with correlation coefficient less than 0.3. Considering the limited value range of elements such as

month, holiday, weekday, etc., I processed them in one-hot format so that they can better fit the training model.

Junfei Sun: In this project, I was mainly responsible for building and implementing the training model, and working with Lingge to complete the preprocessing of the dataset. Based on the concepts I learned, I applied the method of least squares (shortest distance from a point to a line) to training on the dataset and finally established a relationship matrix between the number of bicycles and other influencing factors. In the preprocessing of the data, we call the runtime library that calculates the correlation coefficient to easily view the correlation between the number of bicycles and other factors and use this to decide whether the relevant factors should be deleted.

Jihui Yang: In this project, I was mainly responsible for the analysis and testing of the results, and checked and verified the data set and prediction results and accuracy to ensure that the prediction model was successfully applied to the preprocessed data set

Design ideas

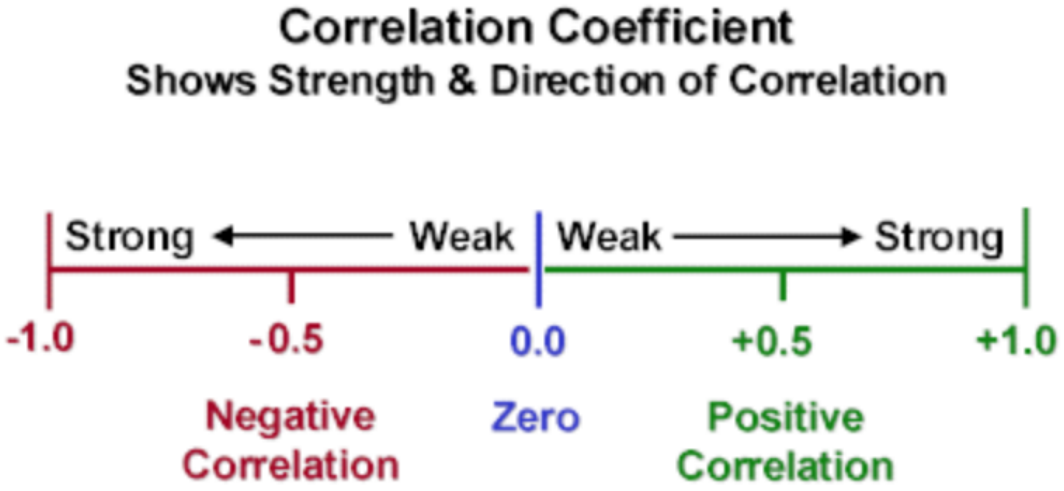
We used the shared bicycle dataset as training set and test set, and predicted the number of shared bicycles rented per day by inputting variables such as date and weather. For the predicted "best travel date", we will determine the best travel date based on the minimum number of rentals for the given dates.

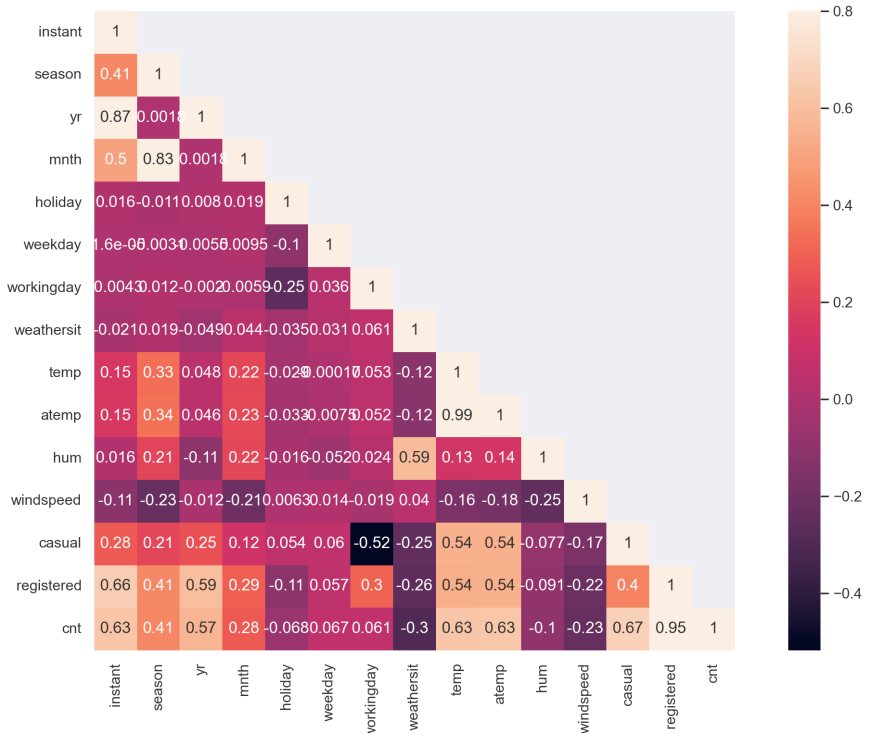
Since this is a linear regression problem, we will use the concept of Linear regression for specific prediction operations.

the programming language is python. Due to the need for large-scale matrix operations, some libraries can greatly improve the running rate, we will use numpy as external libraries here. Also, we plan to use pandas, scatter as an external library due to the need for images to show accuracy.

Design Process

data pre-pocess: The correlation of elements with the target value is analyzed, and the elements with low correlation are removed.





One-hot format processing is done for holiday, weekday, workingday, weathersit.

	season_1	season_2	season_3	...	weekday_4	weekday_5	weekday_6
0	1	0	0	...	0	0	1
1	1	0	0	...	0	0	0
2	1	0	0	...	0	0	0
3	1	0	0	...	0	0	0
4	1	0	0	...	0	0	0
..	...	...	...	...	...	...	...
726	1	0	0	...	1	0	0
727	1	0	0	...	0	1	0
728	1	0	0	...	0	0	1
729	1	0	0	...	0	0	0
730	1	0	0	...	0	0	0

Divide the data set into training set and test set according to 8:2

Build train Model:

By using the least squares method, the model is optimized to obtain the minimum value of the distance from Xi to Y.

$$\frac{\partial(RSS)}{\partial\hat{\beta}} = \frac{\partial(Y^T Y - Y^T X\hat{\beta} - \hat{\beta} X^T Y + \hat{\beta}^T X^T X\hat{\beta})}{\partial\hat{\beta}}$$

$$0 - Y^T X - (X^T Y)^T + 2\hat{\beta}^T X^T X = 0$$

Get the Coefficient of X<->Y

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Get the prediction Y(cnt)

$$\hat{Y} = X\hat{\beta}$$

Verification: verify the testset with the training model

Accuary: calculate the accuracy by the following algorithm,since it is really difficult to make sure the prediction values same with the actual value exactly.

$$R = \left(1 - \frac{testY - testX}{testY}\right) \times 100\%.$$

find $R > 80\%$.

Construct the prediction dataset:

```
forfun=pd.DataFrame({"dteday":['3-17-2022','3-18-2022','3-19-2022','3-20-2022','3-21-2022','3-22-2022','3-23-2022'],
                     "season_1":[1,1,1,1,0,0,0],
                     "season_2":[0,0,0,0,1,1,1],
                     "season_3":[0,0,0,0,0,0,0],
                     "season_4":[0,0,0,0,0,0,0],
                     "mnth_1":[0,0,0,0,0,0,0],
                     "mnth_2":[0,0,0,0,0,0,0],
                     "mnth_3":[1,1,1,1,1,1,1],
                     "mnth_4":[0,0,0,0,0,0,0],
                     "mnth_5":[0,0,0,0,0,0,0],
                     "mnth_6":[0,0,0,0,0,0,0],
                     "mnth_7":[0,0,0,0,0,0,0],
                     "mnth_8":[0,0,0,0,0,0,0],
                     "mnth_9":[0,0,0,0,0,0,0],
                     "mnth_10":[0,0,0,0,0,0,0],
                     "mnth_11":[0,0,0,0,0,0,0],
                     "mnth_12":[0,0,0,0,0,0,0],
                     "weathersit_1":[0,0,0,0,0,1,1],
                     "weathersit_2":[0,1,0,0,0,0,0],
                     "weathersit_3":[1,0,1,1,1,0,0],
                     "weekday_0":[0,0,0,1,0,0,0],
                     "weekday_1":[0,0,0,0,1,0,0],
                     "weekday_2":[0,0,0,0,0,1,0],
                     "weekday_3":[0,0,0,0,0,0,1],
                     "weekday_4":[1,0,0,0,0,0,0],
                     "weekday_5":[0,1,0,0,0,0,0],
                     "weekday_6":[0,0,1,0,0,0,0],
                     "temp": [0.38, 0.38, 0.32, 0.29, 0.34, 0.46, 0.47],
                     "atemp": [0.39, 0.39, 0.35, 0.34, 0.37, 0.45, 0.45],
                     "holiday":[0,0,0,0,0,0,0],
```

Make Prediction: Find the best day suitable for travel. The best day to travel will be the day with the smallest cnt.

Expectation results

In this section, our expectations for this project are:

1. Data set after preprocessing: all the columns of irrelevant factors are successfully eliminated, and only the relevant factors and cnt columns are retained. Some columns of the pruned dataset were successfully converted to "one-hot" format.

2. Successful application of linear regression models

3. When the data set is trained, the accuracy of the test (since the predicted value cannot be guaranteed to be exactly the same as the actual value, an error interval is set here. If the predicted value is within this interval, the prediction is considered successful) above 80%.

4. Prediction: After inputting the relevant data of the next seven days (dteday, season, mnth, holiday, workingday, weathersit, temp, atemp), the program can predict which day is the most suitable for travel.

Actual result

1. Irrelevant columns are successfully proposed, and some related columns are converted to one-hot format. Screenshot below:

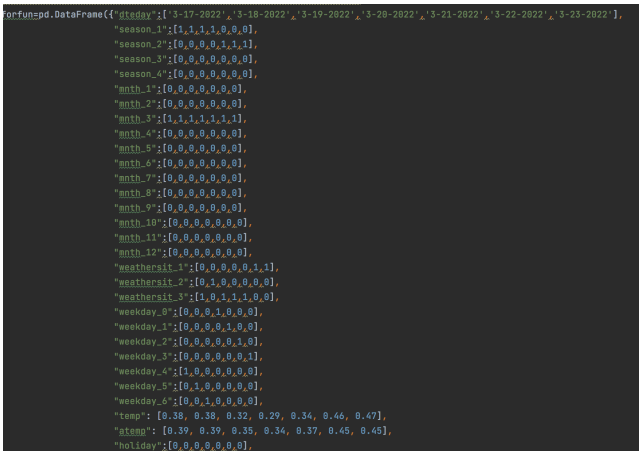
```
     season_1  season_2  season_3  ...  weekday_4  weekday_5  weekday_6
0           1         0         0  ...          0          0          1
1           1         0         0  ...          0          0          0
2           1         0         0  ...          0          0          0
3           1         0         0  ...          0          0          0
4           1         0         0  ...          0          0          0
..        ...       ...       ...  ...        ...        ...        ...
726         1         0         0  ...          1          0          0
727         1         0         0  ...          0          1          0
728         1         0         0  ...          0          0          1
729         1         0         0  ...          0          0          0
730         1         0         0  ...          0          0          0
```

2. Due to technical reasons, the linear model cannot be displayed directly, but the accuracy rate will reflect whether the model has been successfully applied

3. In the test set, the accuracy rate is greater than 80%, and the results are shown in the following figure:

```
[6296. 6880. 6232. 6104. 6408. 6552. 6808. 6664. 6688. 4128. 6608. 5720.
 5752. 6520. 6552. 6632. 6128. 6064. 5624. 6544. 6728. 6600. 6712. 6832.
 6688. 6184. 6616. 7192. 7264. 6488. 7112. 6352. 6552. 6608. 6728. 6728.
 6808. 6888. 6904. 6520. 5944. 6096. 6656. 6680. 6832. 6952. 6944. 6960.
 7304. 7376. 6872. 6856. 7296. 6944. 5960. 4552. 6288. 6400. 6968. 6864.
 5552. 5216. 5856. 6640. 6432. 6480. 6336. 6440. 6024. 6536. 6448. 6136.
 6232. 6624. 6256. 6448. 6776. 6808. 6176. 6184. 6184. 5704. 4104. 5360.
 5536. 5224. 5792. 5040. 5392. 5352. 5464. 4736. 5672. 5776. 5952. 5744.
 5704. 4976. 5392. 5040. 5760. 5688. 5408. 5040. 5264. 5752. 5464. 5904.
 5320. 5144. 5568. 4848. 5424. 5576. 5736. 5120. 4952. 5976. 6168. 5960.
 5400. 5088. 5368. 5056. 5296. 5056. 4896. 5520. 5600. 5776. 4984. 5184.
 5960. 5680. 5112. 3800. 4144. 4024. 3544. 3504. 1800. 3464. 3688. 3616.
 3824. 3360.]
the accuary is: 0.8082191780821918
```

4. The predicted input and output are shown in the following figures




```
[23400. 25040. 23344. 23048. 23816. 26384. 26312.]
The most suitable day for travel is: 3-20-2022
```

Conclusion Discussion

In this assignment, we go through what we've learned about machine learning this semester. A linear regression model was constructed to analyze and predict the usage of shared bicycles. In the end, the expected effect was basically achieved. One of the things that is not very satisfying is that the data forecast for 2022 is a bit higher than our expectations (I guess because the data is for 2011 and 2012, and the 2012 registrations are much higher than 2011 as a whole, so when I calculate In 2022, the results will be magnified.) In addition, the accuracy of the least squares method is not ideal, and finally choose to use sklearn in the python data package to model the data.

The whole experiment is divided into 3 parts：

Code: "main.py", "test_purpose", "bike_data"

Report: ML_GroupProject_Report.pdf

Our dataset: day.csv

Others

I think our linear regression model is very useful to predict the future trend of an event by analyzing the factors that affect it. In my opinion, it can also be used to predict wear and longevity of machine components. It can even be used to predict the price of gold and the direction of stocks. (The premise is to know enough about the factors that affect them. In fact, no one can accurately understand and predict)

For the model we have trained, it may be immature to be honest, there are still many factors that we need to take into account. Including policy, urban planning, and population migration. These will greatly affect the accuracy of the model, making it less accurate in predicting future trends.

Personally, I think that in many cases, the model needs to be constantly updated and adjusted to adapt to new developments, so that the accuracy can be avoided and it can be better suitable for future environments.