BOX OFFICE FORECAST

JINCAI MA

Abstract. Bike-sharing refers to the bicycle sharing services provided by enterprises in campuses, subway stations, bus stations, residential areas, business districts and public service areas. It is a time-sharing model and a new green and environment-friendly sharing economy. In essence, bike-sharing is a new type of transportation rental business – bicycle rental business, which mainly relies on the carrier of (bicycle) bicycles.Can make full use of the city due to rapid economic development caused by the sluggish bicycle travel;Maximize the public road pass rate. The purpose of this project is to predict the demand for bike rental in the D.c. D.C. bike-sharing program by combining historical weather data on bike-sharing usage patterns.

Contents

Date: (None).
1991 *Mathematics Subject Classification.* Forecast use of a city bikeshare system.

1. Introduction

(1)Use recipe ingredients to categorize the cuisine. Given the name of the condiment, predict the cuisine to which the dish belongs.
(2)The data comes from Kaggle https://www.kaggle.com/c/bike-sharing-demand.
(3) In the dataset,including the recipe ID, the dish, and the list of ingredients for each recipe (variable length).The data is stored in JSON format.
1.train.json- A training set that contains the recipe ID, dish type, and ingredient list 2.test.json- A test set containing a recipe ID and a list of ingredients 3.sample_submission.csv-Properly formatted sample submission document

2. Data Analysis

First of all, our work can be divided into the following steps:
(1) Data Import And Introduction:
a.Import the JSON file with Pandas: We can get the data set of dish names, including 39774 training data and 9944 test samples.To see the distribution of our data set and the total variety of dishes, we printed out some of the data samples.

	id	cuisine	ingredients
0	10259	greek	[romaine lettuce, black olives, grape tomatoes, garlic, pepper, purple onion, seasoning, garbanzo beans, feta cheese crumbles]
1	25693	southern_us	[plain flour, ground pepper, salt, tomatoes, ground black pepper, thyme, eggs, green tomatoes, yellow corn meal, milk, vegetable oil]
2	20130	filipino	[eggs, pepper, salt, mayonaise, cooking oil, green chilies, grilled chicken breasts, garlic powder, yellow onion, soy sauce, butter, chicken livers]
3	22213	indian	[water, vegetable oil, wheat, salt]
4	13162	indian	[black pepper, shallots, cornflour, cayenne pepper, onions, garlic paste, milk, butter, salt, lemon juice, water, chili powder, passata, oil, grou...
5	6602	jamaican	[plain flour, sugar, butter, eggs, fresh ginger root, salt, ground cinnamon, milk, vanilla extract, ground ginger, powdered sugar, baking powder]

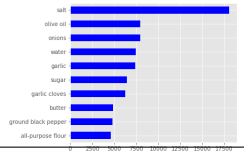
b.Total dish classification

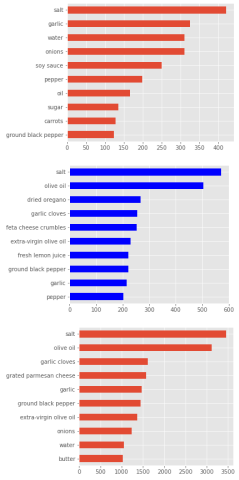
There are 20 dishes in total, which are: ['brazilian' 'british' 'cajun_creole' 'chinese' 'filipino' 'french' 'greek' 'indian' 'irish' 'italian' 'jamaican' 'japanese' 'korean' 'mexican' 'moroccan' 'russian' 'southern_us' 'spanish' 'thai' 'vietnamese']

(2) a.The data set is divided into Features and Target Variables.
b.Features:'ingredients', we were given the names of the ingredients contained in each dish; Target variable:'cuisine', is the classification of cuisines that we want to predict.
c.Extract the Feature of training data set into train_integredients variable Extract the Target Variables into the train_Targets variable.

```
0        [romaine lettuce, black olives, grape tomatoes, garlic, pepper, purple onion, seasoning, garbanzo beans, feta cheese...
1        [plain flour, ground pepper, salt, tomatoes, ground black pepper, thyme, eggs, green tomatoes, yellow corn meal, mil...
2        [eggs, pepper, salt, mayonaise, cooking oil, green chilies, grilled chicken breasts, garlic powder, yellow onion, so...
3                                                  [water, vegetable oil, wheat, salt]
4        [black pepper, shallots, cornflour, cayenne pepper, onions, garlic paste, milk, butter, salt, lemon juice, water, ch...
                                                ...
39769    [light brown sugar, granulated sugar, butter, warm water, large eggs, all-purpose flour, whole wheat flour, cooking ...
39770    [KRAFT Zesty Italian Dressing, purple onion, broccoli florets, rotini, pitted black olives, Kraft Grated Parmesan Ch...
39771    [eggs, citrus fruit, raisins, sourdough starter, flour, hot tea, sugar, ground nutmeg, salt, ground cinnamon, milk, ...
39772    [boneless chicken skinless thigh, minced garlic, steamed white rice, baking powder, corn starch, dark soy sauce, kos...
39773    [green chile, jalapeno chilies, onions, ground black pepper, salt, chopped cilantro fresh, green bell pepper, garlic...
Name: ingredients, Length: 39774, dtype: object
0            greek
1        southern_us
2           filipino
3             indian
4             indian
            ...
39769         irish
39770       italian
39771         irish
39772       chinese
39773       mexican
Name: cuisine, Length: 39774, dtype: object
```

(3) Data Visualization:What are the top 10 most frequently used ingredients? What are the 10 most common ingredients in filipino,greek and Italian cuisine?





(4) Data Cleaning:Since dishes contain a large number of ingredients, and since the same ingredients can vary in numbers, tenses, and so on, we considered sifting through a potatos to remove any such differences.

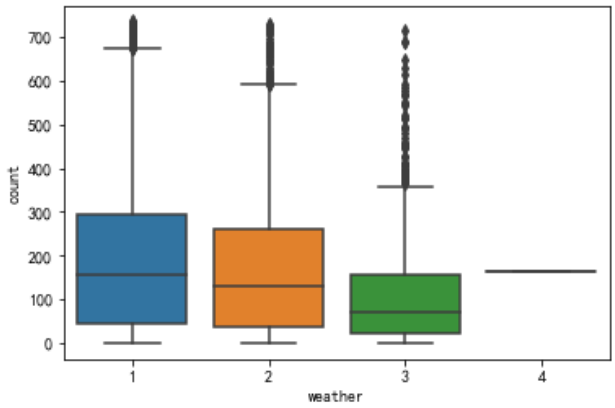
```
处理训练集...
菜品佐料：
['chopped tomatoes', 'fresh basil', 'garlic', 'extra-virgin olive oil', 'kosher salt', 'flat leaf parsley']
去除标点符号之后的结果：
['chopped tomatoes', 'fresh basil', 'garlic', 'extra virgin olive oil', 'kosher salt', 'flat leaf parsley']
去除时态和单复数之后的结果：
chopped tomato fresh basil garlic extra virgin olive oil kosher salt flat leaf parsley
处理测试集...
菜品佐料：
['eggs', 'cherries', 'dates', 'dark muscovado sugar', 'ground cinnamon', 'mixed spice', 'cake', 'vanilla extract', 'self raising flour',
'sultana', 'rum', 'raisins', 'prunes', 'glace cherries', 'butter', 'port']
去除标点符号之后的结果：
['eggs', 'cherries', 'dates', 'dark muscovado sugar', 'ground cinnamon', 'mixed spice', 'cake', 'vanilla extract', 'self raising flour',
'sultana', 'rum', 'raisins', 'prunes', 'glace cherries', 'butter', 'port']
去除时态和单复数之后的结果：
egg cherry date dark muscovado sugar ground cinnamon mixed spice cake vanilla extract self raising flour sultana rum raisin prune glace
cherry butter port
```

(5) Feature extraction:

a.We convert the ingredients of the dish into a numerical feature vector.Consider that most dishes include salt, water, sugar, butter, etc,We will consider weighting the seasonings according to the occurrence times of the seasonings, that is, the more the occurrence times of the condiments, the lower the discriminability of the condiments.The feature we adopt is TF-IDF.

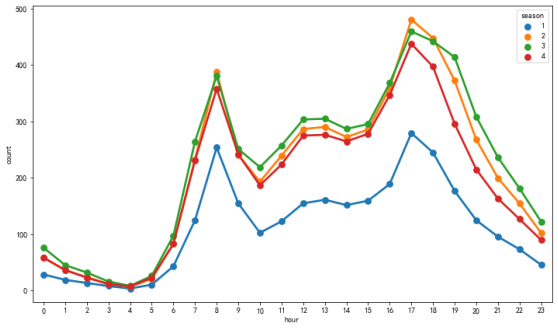
b.We can get the characteristics:['greek','southern_us','filipino','indian','indian', 'jamaican','spanish','italian','mexican','italian']

(6) The impact of weather

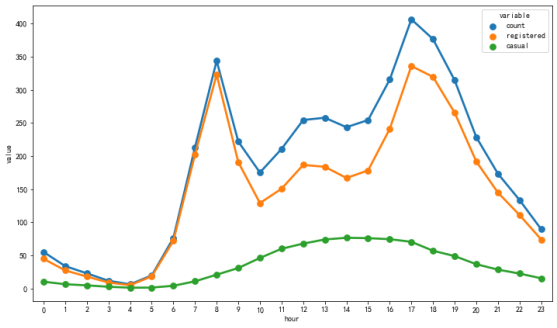


(7) Impact of season,week,registered and non-registered users on cycling usage trends

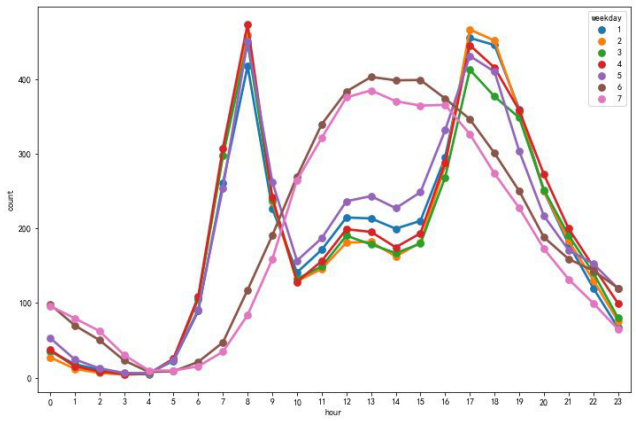
a.For different times of the day, there is a clear trend in the use of Shared bikes, with two distinct peaks, in line with people's understanding of morning peak and evening peak.The trends were the same for all four seasons, except that usage in spring was slightly lower than in the other three.



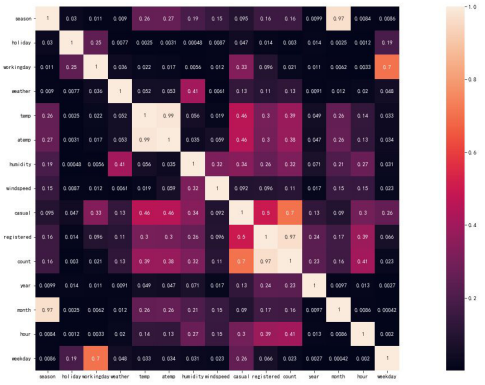
b.The usage of registered users accounts for the majority of the total usage, and the trend is consistent with the total usage trend, rather than that of registered users. The usage at different times of the day does not change much, and the trend is similar to the usage trend at weekends.



c.From Monday to Friday, there are two peak usage periods, while on weekends, the usage trend is completely different from that on weekdays. The usage trend changes from bimodal to flat unimodal, and the peak usage period is concentrated at 11-17 o 'clock.



(8) Draw the thermal diagram of the correlation coefficient



cost

3. Build Model

1.Separate the training set and test set.
2Remove unwanted eigenvalues:'casual','count','datetime','registered','date','atemp','month','year','season','weat
3.Cross validation is used to determine the optimal parameters.
4.View the selected optimal parameters:'max_depth': 20, 'n_estimators': 150
5.Apply the optimal parameters to the model, it can be obtained
Accuracy on test set: 0.6945996275605214
Recall rate on test set: 0.7379725915789399

4. Conclusions

Through this Kaggle project, I practiced by myself to have a deeper understanding of data visualization and to explore the structure and rules of data by means of drawing and tabulating.

List of Todos

(A. 1) School of Computer Science,, Xi'an Shiyou University, Shaanxi 710065, China
Email address, A. 1: xxx@tulip.academy