**Regression Part I: Linear regression model**

After we import the data from ‘dc’ bike sharing from 2011 to 2012, we will create dummy variables for categorial type variable 'weather' and 'season'. Since for categorial type variable the symbolized number does not represent higher value from 1 to 4. Also for the ‘hour’, we will use the C() function in the formula to automatically create the dummy instead of input x one by one.

For the 'season' variable, we will use temperate as dependent variable to compare temperature for each season using s\_4 as the benchmark. And the result is rank of average temperature for four seasons from low to high is : s\_1, s\_4, s\_2, s\_3.

图片包含 文字, 屏幕截图

描述已自动生成

Then for the linear regression model, we will use all variables showing in the data set and 'cnt' (the total number of people rent the sharing bike) as the dependent variable and it will be the same for the regression tree in the later part. The result is showing as below:

We can see that the highest number of people using share bike is around 17:00 pm which is the time that most people will go home from work. And for X0 to X5, it separately represents 'holiday', 'workingday', 'TF', 'TFF', 'Humidity', 'WindSpeed'. We can see that more people choose to use sharing bike during work day than holiday and the higher the temperate with lower humidity and windspeed, more people would like to use shared bike as commute method.

图片包含 文字

描述已自动生成图片包含 文字, 黑板, 记分牌

描述已自动生成

Also, in order to have deeper understanding of our data for casual user and registered user we have separated the season and temperature effect, working day and holiday to avoid multicollinearity. The following are the four models we build:

图片包含 文字, 屏幕截图

描述已自动生成 # 1. ('holiday', 'TF', 'Humidity')

For casual users, We can see from the result that on holiday there are more casual users renting bike. Also, the higher temperature and lower humidity, more casual users will rent the bike.

图片包含 文字

描述已自动生成 # 2. ('holiday', 'Humidity', 's\_1', 's\_2', 's\_3')

More casual users rent bike on holidays when humidity is low and in season that has higher rank in temperature

图片包含 文字

描述已自动生成 # 3. ('workingday', 'TF', 'Humidity')

More casual users are less likely to rent bike on workingday and if they do more casual users rent bike when temperature is high and humidity is low

图片包含 文字, 记分牌

描述已自动生成 # 4. ('workingday', 'Humidity', 's\_1', 's\_2', 's\_3')

More casual users are less likely to rent bike on workingday and if they do more casual users rent bike when humidity is low and season have higher rank in temperature

For registered users, We can see from the result that on holiday there are less registered users renting bike. Since for the registered user, the main reason they register is that they will use a lot and we can imagine those people are because of work commute need. In this case, when it comes to holiday, those people will just rest and not using bikes. Also, the higher temperature, more casual users will rent the bike.

图片包含 文字, 屏幕截图

描述已自动生成 # 1. ('holiday', 'TF', 'Humidity')

Less registered users rent bike on holidays and if they do, more people will rent bike when temperature is high and humidity is low

图片包含 文字

描述已自动生成 # 2. ('holiday', 'Humidity', 's\_1', 's\_2', 's\_3')

Less registered users rent bike on holidays and if they do, more people will rent bike when humidity is low. the rank of the number of people renting bikes for each season from low to high is s\_1, s\_2, s\_4, s\_3. This is different from the result we get from 'TF' (previously we have the temperature order from low to high as s\_1, s\_4, s\_2, s\_3)

图片包含 屏幕截图

描述已自动生成 # 3. ('workingday', 'TF', 'Humidity')

More registered users are renting bike on workingday when humidity is low and seasons have higher rank in temperature

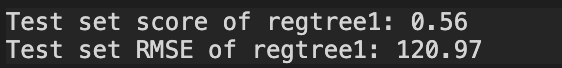
图片包含 文字

描述已自动生成 # 4. ('workingday', 'Humidity', 's\_1', 's\_2', 's\_3')

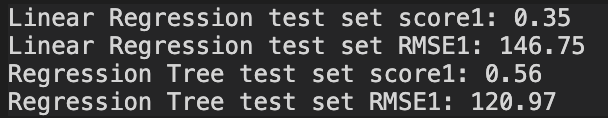
More registered users are renting bike on workingday when humidity is low. And the rank of the number of people renting bikes for each season from low to high is s\_1, s\_2, s\_4, s\_3. This is different from the result we get from 'TF' (previously we have the temperature order from low to high as s\_1, s\_4, s\_2, s\_3)

**Regression Part II: Regression tree model**

We will use the same variables in the OLS model for comparison and split the dc dataset into 80% train, 20% test. The following is the result we have for instantiating a Decision Tree Regressor.



Then we will compare the performance with OLS, and we will use Root Mean Square Error to see which model has a better fit. we can see from the result that regression tree has lower RMSE and higher score which means that it has better model fit. The unit of RMSE is same as dependent variable. If your data has a range of 0 to 100000 then RMSE value of 3000 is small, but if the range goes from 0 to 1, it is pretty huge.



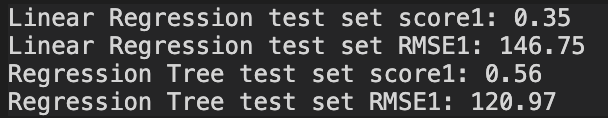
In this case, we have RMSE around 130 and the range of casual users is from 0 to 977, it is about 0.1 of the largest number and the R-squared is 0.6, which shows that it is also a good fit for the total user

Also, we will evaluate the list of Mean Square Error obtained by 10-fold CV, we can see from the result that training set and test set have similar amount of RMSE which means that the model we built is a good one. And the regression tree is better than the linear regression.

图片包含 瓶子

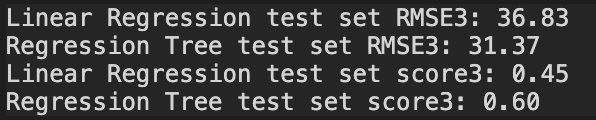
描述已自动生成

In order to have deeper valuation of our model, we will separate the situation for working days and holidays as we did in the liner regression model, and the dependent variable will be ‘cnt’(the total number of people rent the sharing bike) The result are the similar as shown above for the day only in working day.

图片包含 瓶子

描述已自动生成

Then we will use ‘casual’ as dependent variable for another valuation. And the result is showing below:

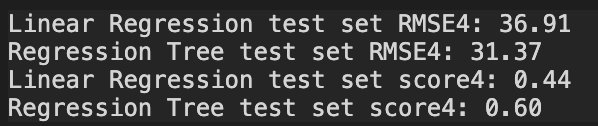
 图片包含 瓶子

描述已自动生成

In this case, we have RMSE around 30 and the range of casual users is from 0 to 367, it is about 0.1 of the largest number and the R-squared is 0.3, which shows that it is also a good fit for the casual user. And the regression tree is better than the linear regression.



In order to have deeper valuation of our model, we will separate the situation for working days and holidays as we did in the liner regression model, and the dependent variable will be ‘casual’ The result are similar as shown above for the day only in working day.

图片包含 瓶子

描述已自动生成

**Prediction:**

After we compared our models with two regression, we will use the web scrapping data from 2017 weather and bike sharing to see how our model performs. And we use original dc data as train set and dc 2017 data as test set. The variables will be ‘TF’, ‘humidity, ‘windspeed’, and ‘hour’, and the dependent variable is ‘cnt’. We can see from the result that the linear regression model is better than the regression tree model instead and the RMSE has increased a lot from around 120 to around 440. And the training set RMSE are much lower than the test set RMSE which means that we have over-fitting the data. And the reason is because that we do not have enough variables as we did in the previous data set of dc (which contains over 10 different variables). If we do not have the limitation of online sourcing data set of web scrapping and have more variables, we can have better prediction

