Course > Final Exam > Final Exam > Predicting Demand

Audit Access Expires Aug. 12, 2019
You lose all access to this course, including your progress, on Aug. 12, 2019.

Predicting Demand

Bike-sharing systems are appearing all over the world; examples include Citi Bike in New York City, Santander Cycles in London, and ofo in China. These services allow users to make short-term bike rentals. In *docked systems*, docking stations are set up in prespecified locations, and users must pick up and return the bike to a docking station within the system. In *dockless systems*, users are able to pick up and return bikes to any desired location (pickups pending availability). There is a lot of research in Bike-sharing systems and as a start, in this problem, we will attempt to understand the factors that influence a high demand for this service.

Dataset: bikes.csv

In the dataset above, each observation represents one hour of the day (10886 hours). Here is a detailed description of the variables:

- **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
- **count**: number of total rentals
- **demand_level**:1 if **count** is at least 250, 0 otherwise
- **hour**: the hour of the day (0-23)

In this problem, we will use various classification methods to try to predict the demand level.

Problem 1 - Exploratory Data Analysis

0.0/3.0 points (graded)

Which season has the most rentals?

○ Summer

○ Fall ✓

○ Winter

○ Spring

What is the average temperature in Celsius?

Answer: 20.23086

What is the average temperature in Celsius during the high demand hours?

High demand is defined by demand_level = 1.

Answer: 24.48587

Submit You have used 0 of 2 attempts

Answers are displayed within the problem

Problem 2.1 - Preparing the Data

0.0/3.0 points (graded)

We will now split the data into a training and testing set. To do this, we use the sample.split() function. Which variable will be used in this function?

○ temp

○ count

○ demand_level ✓

○ season

Set your random seed to 100 and create a training and test set using the sample.split() function in the caTools library, with 70% of the observations in the training set and 30% in the testing set.

Why do we use the sample.split() function?

○ It is the most convenient way to randomly split the data

○ It balances the independent variables between the training and testing sets

○ It balances the dependent variable between the training and testing sets ✓

How many observation are there in the training set?

Answer: 7620

Submit You have used 0 of 2 attempts

Answers are displayed within the problem

Problem 2.2 - Simple Logistic Regression

0.0/2.0 points (graded)

Train a logistic regression model using temp as the independent variable. What is the coefficient of temp?

Answer: 0.110214

Submit You have used 0 of 2 attempts

ℹ Answers are displayed within the problem

Problem 2.3 - Simple Logistic Regression

0.0/5.0 points (graded)

Using your logistic regression model, obtain predictions on the test set. Then, using a probability threshold of 0.5, create a confusion matrix for the test set. What is the (test) accuracy of your logistic regression model?

Answer: 0.72902633

Our baseline model in classification is to always predict the most frequent outcome in the test set. What is the (test) accuracy of this baseline model?

Answer: 0.70024495

What is the true positive rate of your logistic regression model?

Answer: 0.27783453

What is the false positive rate of your logistic regression model?

Answer: 0.07783122

Currently, we are predicting many more low demand observations than high demand observations. Which of the following is a way to change that?

○ It is impossible to predict more high demand with this model. To change these results, another model can be used.

○ To predict more high demand, decrease the prediction threshold. ✓

○ To predict more high demand hours, increase the prediction threshold.

○ To predict more high demand hours, create more observations with high demand.

Submit   You have used 0 of 2 attempts

ℹ Answers are displayed within the problem

Problem 3.1 - Adding More Variables

0.0/2.0 points (graded)

We would now like to train a logistic regression model using all of the variables in the training set. Which of the following is true?

□ Weather and temp are highly correlated.

□ Season and weather are highly correlated.

□ Workingday and holiday are not highly correlated. ✓

□ Temp and atemp are highly correlated. ✓

Train a logistic regression model now using all of the following variables in the training set:

season, holiday, workingday, weather, temp, humidity, windspeed, and hour

Which of the following variables are significant at a level of 0.001 or less?

□ season ✓

□ holiday

□ workingday

□ weather

□ temp ✓

□ humidity ✓

□ windspeed

□ hour ✓

Submit You have used 0 of 2 attempts

Answers are displayed within the problem

Problem 3.2 - Adding More Variables

0.0/5.0 points (graded)

Using your new logistic regression model, obtain predictions on the test set. Then, using a probability threshold of 0.5, create a confusion matrix for the test set.

What is the (test) accuracy of your logistic regression model?

Answer: 0.7672994

Which of the following is true?

□ Close to a third of time that there is high demand, the model will predict high demand.

□ Almost half of the times that there is high demand, the model will predict high demand. ✓

□ About 75% of the times that there is high demand, the model will predict high demand.

□ About 10% of the times that there is low demand, the model will predict high demand. ✓

□ About 25% of the times that there is low demand, the model will predict high demand.

□ About 7% of the times that there is low demand, the model will predict high demand.

Plot the ROC curve for your logistic regression model. Which logistic regression threshold is associated with the lower-left corner of the ROC plot (true positive rate 0 and false positive rate 0)?

○ 0

○ 0.5

○ 1 ✓

At roughly which logistic regression cutoff does the model achieve a true positive rate of 80% and a false positive rate of 40%?

○ 0.01

○ 0.19 ✓

○ 0.37

○ 0.55

○ 0.73

○ 0.91

What is the AUC for your logistic regression model?

Answer: 0.8031658

Submit

You have used 0 of 2 attempts

Answers are displayed within the problem

Problem 4.1 - CART

0.0/4.0 points (graded)

Set the random seed to 100.

Then use the caret package and the train function to perform 10-fold cross validation with the training data set to select the best cp value for a CART model that predicts the dependent variable demand_level using all of the possible independent variables except count which was used to define the dependent variable. Select the cp value from a grid consisting of the values 0.0001, 0.0002, 0.0003, ..., 0.02.

Remember to convert the demand_level column to a factor variable.

If you have called your training set train, use the following code:

train$demand_level = as.factor(train$demand_level)

Which cp value maximizes the cross-validation accuracy?

Answer: 0.001

If you would like to view the tree, export it as a PDF from RStudio.

What does the first split indicate? (2 points)

○ There will be a high demand of bikes before 7 AM.

○ There will not be a high demand of bikes before 7 AM. ✓

○ If the hour is before 7 AM, we should look at the temperature.

○ If the hour is before 7 AM and the temperature is less than 17, there will not be a high demand

Submit You have used 0 of 2 attempts

Answers are displayed within the problem

Problem 4.2 - CART

0.0/2.0 points (graded)

What is the (test) accuracy of your CART model?

Answer: 0.88763013

What does the CART model predict on a Saturday, spring day at 9 AM when the temperature is 15 degrees Celsius?

○ high demand

○ low demand ✓

○ Not enough information

Submit You have used 0 of 2 attempts

Answers are displayed within the problem