
Do we need more bikes?

Project in Statistical Machine Learning

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 In this project we develop, and study different statistical machine learning models
2 for predicting whether the number of available bikes at a given hour should be
3 increased, a project by the District Department of Transportation in Washington
4 D.C. The training data set consists of 1600 instances of hourly bike rentals, and
5 a test set of 400 instances. The models for prediction we have used are: *Logistic*
6 *regression*, *Discriminant methods: LDA, QDA, k- Nearest Neighbour*, and *Tree*
7 *Based Methods*. We have found that THE MODEL gives best prediction, with
8 accuracy ??????

9 **1 Plan**

10 **1.1 From Intro**

11 (i) Explore and preprocess data

12 (ii) try some or all classification methods, which are these?

13 • Logistic Regression

14 • Discriminant analysis: LDA, QDA

15 • K-nearest neighbor

16 • Tree-based methods: classification trees, random forests, bagging

17 • Boosting

18 (iii) Which of these are to be "put in production"?

19 **1.2 From Data analysis task**

20 • Can any trend be seen comparing different hours, weeks, months?

21 • Is there any difference between weekdays and holidays?

22 • Is there any trend depending on the weather?

23 **1.3 From Implementation of methods**

24 Each group member should implement one family each, who did what shall be clear!

25 DNNs are encouraged to be implemented, do this if there is time. (DNN is not a thing a group

26 member can claim as their family.)

27 Implement a naive version, let's do: *Always low_bike_demand*

28 **1.3.1 What to do with each method**

29 1. Implement the method (each person individually)

30 2. Tune hyper-parameters, discuss how this is done (each person individually)

31 3. Evaluate with for example cross-validation. Don't use E_{k-fold} (what is that?) (need to do

32 together)

33 4. (optional) Think about input features, are all relevant? (together)

34 Before training, unify pre-processing FOR ALL METHODS and choose ONE OR MULTIPLE

35 metrics to evaluate the model. (is it necessary to have the same for all?, is it beneficial?) Examples:

36 • accuracy

37 • f1-score

38 • recall

39 • precision

40 Use same test-train split for ALL MODELS

41 2 Theoretical Background

42 2.1 Mathematical Overview of the Models

43 2.1.1 Logistic Regression

44 The backbone of logistic regression is linear regression, i.e. finding the least-squares solution to an
45 equation system

$$X\theta = y \quad (1)$$

46 given by the normal equations

$$X^T X \theta = X^T y \quad (2)$$

47 where X is the training data matrix, θ is the coefficient vector and b is the training output. The
48 parameter vector is then used in the sigmoid function:

$$\sigma(z) = \frac{e^z}{1 + e^z} : \mathbb{R} \rightarrow [0, 1], \quad (3)$$

$$z = x^T \theta, \quad (4)$$

49 where x is the testing input. This gives a statistical interpretation of the input vector. In the case of a
50 binary True/False classification, the value of the sigmoid function then determines the class.

51 2.2 Input Data Modification

52 By plotting the data and analyzing the .csv file, some observations were made. The different inputs
53 were then changed accordingly:

- 54 • *Kept as-is:* weekday, windspeed, visibility, temp
- 55 • *Modified:*
 - 56 – month - split into two inputs, one cosine and one sine part. This make the new inputs
 - 57 linear and can follow the fluctuations of the year. The original input was discarded.
 - 58 – hour_of_day - split into three boolean variables: demand_day, demand_evening,
 - 59 and demand_night, reflecting if the time was between 08-14, 15-19 or 20-07 respec-
 - 60 tively. This was done because plotting the data showed three different plateaues of
 - 61 demand for the different time intervals. The original input was discarded.
 - 62 – snowdepth, precip were transformed into booleans, reflecting if it was raining or
 - 63 if there was snow on the ground or not. This was done as there was no times where
 - 64 demand was high when it was raining or when there was snow on the ground.
- 65 • *Removed:* cloudcover, day_of_week, snow, dew, holiday, summertime. These were
- 66 removed due to being redundant (e.g. summertime), not showing a clear trend (e.g.
- 67 cloudcover), giving a worse score when used, or all three (e.g. day_of_week).

3 Data Analysis

In the data, there are some numerical and categorical features:

- *Numerical*: temp, dew, humidity, precip, snow, snowdepth, windspeed, cloudcover and visibility.
- *Categorical*: hour_of_day, day_of_week, month, holiday, weekday, summertime, and increase_stock

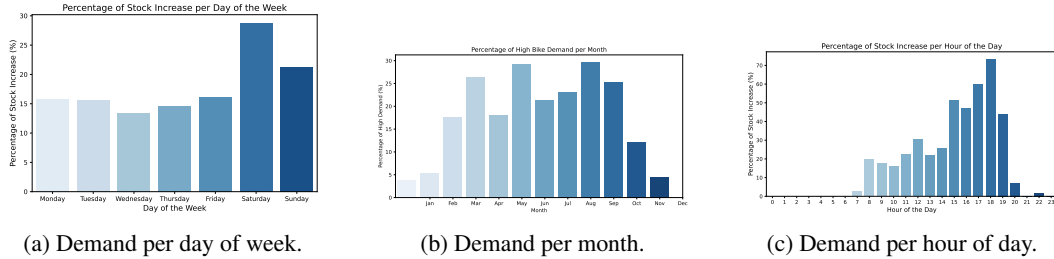


Figure 1: Bike demand vs. day of week and month.

Figure 2: Bike demand vs. temperature LÄGG IN FLER HÄR TYP.

There are some trends seen in the data when it comes to time and weather. From figure 1, one can see a periodic relationship for the months, where there is a higher demand during the warmer months, loosely following a trigonometric curve. Over the week, the demand is rather stable, with a peak on the weekend, especially Saturdays. Looking at the weather; if there is rain or if there is snow on the ground, there is close to always low demand. Cloudcover did not make a big impact, which is also intuitive, as a cloudy day does not make biking more difficult. Dew point also does not have a clear trend, while humidity however has a clear trend downwards as the humidity increases.

The overall trend is that about one eighth of observations correspond to a high bike demand. During the night, or in bad weather, the demand is (intuitively) low. But during rush hour (fig. 2), the demand is very high, and should probably be increased in order to minimize excessive CO₂ emissions.