

---

# 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 Logistic Regression

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

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

45 given by the normal equations

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

46 where  $X$  is the training data matrix,  $\theta$  is the coefficient vector and  $b$  is the training output. The  
47 parameter vector is then used in the sigmoid function:

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

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

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