UNIVERSITY OF SOUTHERN DENMARK
FACULTY OF BUSINESS AND SOCIAL SCIENCES

Topics in Econometrics

Winter 2020-21

1. Assigner: Christian Møller Dahl and Jørgen T. Lauridsen
2. Hand-out: December 21, 2020, at 12:00 o'clock
3. Hand-in: January 18, 2021, at 12:00 o'clock
4. All pages, incl. the front page, should include the following: Name and birth date
5. All pages must be numbered.

Form of examination for the certificate:
1) Take-home assignment and 2) an oral exam.

Supplemental information for the form of examination:

Ad 1) Report written in groups of 3-4 students (groups that are formed by the course instructors). Smaller groups of 1-2 students require permission from the course instructors.
Duration: Deadline for handing in will appear from the examination plan.
Location: Home assignment.
Internet Access: Necessary.
Hand out: Digital Exam.
Hand in: Digital Exam.
Extent: 30-40 standard pages.
Exam Aids: All exam aids allowed.

Ad 2) 20 minutes are set aside for the oral exam.

The examination takes its starting point in the report and a presentation hereof. The examination also includes supplementary questions on theoretical, methodological and, if relevant, practical topics associated with the submitted report.

Grading according to the Danish 7-point scale. The grading is an overall assessment of the report and the performance at the oral exam. Grading based on the performance of the individual student compared to the learning goals.

Exam question:

Buying a car is never an easy task, especially when buying a used one. So many different factors go into determining the price of a vehicle that it's difficult to accurately predict what one should be paying. However, you are up for a challenge! The fundamental question you seek to answer can be summarized as follows: "Is it possible to predict the price of a used car based on **craigslist** (hosted by www.kaggle.com and available from "Used Cars Dataset: Vehicles listings from Craigslist.org" https://www.kaggle.com/austinreese/craigslist-carstrucks-data)

Problem: Based on the dataset described on "Used Cars Dataset: Vehicles listings from Craigslist.org" you are interested in accurately predicting:

1. The sales price of a used car at a given time and place (consider only times and places spanned by the locations and the time periods available in the data set).
2. If the sales price of a used car is much higher than the median sales price. Please consider alternative measures of "**much** higher sales prices" and motivate your favorite choice. (Hint: Consider this as a classification problem).

Use statistical learning techniques, and as many data sources as possible to obtain the most accurate predictions under **1**. and **2**. In particular, under **1**. and **2.**, search for and identify all the important/relevant features (or combinations of features and/or engineered features) as well as the best possible prediction method/model. In your search for the best possible prediction methods/models please consider tree based methods, shrinkage methods as well as regression models (in all its many forms), logistic models and linear discriminant analysis. Be sure to fit the models on a training set (which you have defined properly) and to evaluate/compare performance on a test set (which you have defined properly). Please, also consider as many factors/predictors as possible.

In relation to **2.** please explicitly consider the following questions:

Q1. Define which variables to use as historical information and investigate to which extent data reduction – in particular (but not necessarily exclusively) in the form of principal component and/or factor analysis – applies for a better understanding of data, and to which extent data reduction may apply for further utilization of data for a discriminant function.

Q2. Suggest a discriminant function where historical information is used to predict sales prices that are much higher than the median sales price.

Q3. Use the discriminant function to used car sales prices and discuss the adequacy and applicability of the model for classification.

Q4. Use method(s) from the course for supplemental analyses that shed additional light on the questions **Q1-Q3**.

Present, discuss and motivate your preferred prediction models under both **1**. and **2**. and conclude by addressing the following questions:

Q5. Based on the preferred prediction model under **1**. what can you infer about the role and importance of the available data in predicting used car sales prices?

Q6. Based on the preferred prediction model under **2**. what can you infer about the role and importance of the available data in predicting sales prices of a used cars that are much higher than the median sales price?