REVIEWER‘S OPINION OF FINAL THESIS

I. IDENTIFICATION DATA

Thesis name:	**Pricing and data: long-distance bus routes**
Author’s name:	**Mohammad Asad Ali**
Type of thesis :	Master’s thesis
Faculty/Institute:	Faculty of electrical engineering
Department:	Department of Computer Science
Thesis reviewer:	Ivan Nikolaev
Reviewer’s department:	External - Barclays Investment Bank

II. EVALUATION OF INDIVIDUAL CRITERIA

Assignment **A**

Evaluation of thesis difficulty of assignment.

In this work the author tackles prediction of ticket prices for inter-city bus routes. This is a challenging problem. The prices depend on many factors, such as seasonality, customer demand and competition between companies, among others. There is no recent publicly available datasets, so the data needs to be collected from the web.

Satisfaction of assignment **D**

Assess that handed thesis meets assignment. Present points of assignment that fell short or were extended. Try to assess importance, impact or cause of each shortcoming.

Real bus pricing data was collected and used to create a dataset. This dataset was used to train and evaluate different regression models. Unfortunately, the collected data was limited to a very small number of bus routes. The main problem is in evaluation, where there is no proper separation of training and testing data and so the evaluation results do not provide a meaningful conclusion, apart from highlighting model overfitting.

Method of conception **B**

Assess that student has chosen correct approach or solution methods.

The student has created a viable pipeline for solving the problem. Real bus pricing data was collected from the website of two different bus companies over a period of time. That data was used to create a dataset and to train and validate different regression models.

Technical level **C**

Assess level of thesis specialty, use of knowledge gained by study and by expert literature, use of sources and data gained by experience.

The student has created a dataset and used to train different regression models on it, most of them based on decision tree methods.

Formal and language level, scope of thesis **B**

Assess correctness of usage of formal notation. Assess typographical and language arrangement of thesis.

The work is separated into logical chapters, proper use of references is employed. There are some typographical and grammatical errors and some of the theoretical explanations are incoherent, but overall the quality is good.

Selection of sources, citation correctness **D**

Present your opinion to student’s activity when obtaining and using study materials for thesis creation. Characterize selection of sources. Assess that student used all relevant sources. Verify that all used elements are correctly distinguished from own results and thoughts. Assess that citation ethics has not been breached and that all bibliographic citations are complete and in accordance with citation convention and standards.

Starting with the state-of-the-art overview, there is little information provided. The author mentions a few works in the area, but only goes on to say that they used outdated methods. He does not discuss or explain the methods used or the results that were achieved.

In the same section the author gives some theoretical background on different machine learning techniques which sometimes lacks coherence. At the end, definitions of the evaluation metrics used are given.

Additional commentary and evaluation

Present your opinion to achieved primary goals of thesis, e.g. level of theoretical results, level and functionality of technical or software conception, publication performance, experimental dexterity etc.

The data pipeline chapter misses a lot of crucial information: which fields were scraped, how often, for how long, what parameters were used (one-way vs return, number of passengers, etc). From this chapter, it is not very clear what data was collected and what features were constructed. Also, for some features it is hard to tell how they were calculated, i.e. number of free seats.

In the dataset chapter, definitions for sample, connection and route is given. Statistics are provided, which show an extremely large number of samples per very small number of connections. The number of routes is not given, but it has to be smaller than the number of connections by definition. Later during evaluation, cross-validation is performed on the level of samples. Given the average number of samples per connection, this effectively means that the training and testing folds will have exactly same distributions. This is confirmed in the results where training and testing errors are the same for all the methods. This could mean that the models could be highly overfitted with no way to tell. A more fair comparison would be sampling training and testing dataset on the level of connections.

Analysis of the dataset and feature importance is provided, but given potential problems with overfitting, it is hard to draw meaningful conclusions from the results.

III. OVERALL EVALUATION, QUESTIONS FOR DEFENSE, CLASSIFICATION SUGGESTION

Summarize thesis aspects that swayed your final evaluation. Please present apt questions which student should answer during defense.

Given the flaw in the evaluation methodology, as well as the overall quality of the work, the proposed grade is **D**. If the author is able to provide a revised evaluation which addresses the overfitting problems, then the proposed grade is higher.

I evaluate handed thesis with classification grade **D**

Date: **16.06.2020**

Signature: Ivan Nikolaev