# Final Report

Novikov Vitalii, Pozdena Nicolas, Prinz Paul, Shapovalov Anton, Wadhwani Amar

## Preface

This report was created as part of the Maters study program Data Science of the University of Applied Sciences Vienna. All authors contributed equally on this project and have been part from the beginning and were part of all decision made.

All of the used data is openly accessible and the sources are referenced in the chapters below.

We would like to thank Dr. Andreas Reschreiter for his advisory role on the report.

## Introduction

The aim of this report was to describe and document the process of building a fare prediction model for chicago taxi fares, based on the openly available data from the city of chicago (dataset). The reason for this project was to train and gain experience on machine learning model, handling large datasets, and planning and finishing a datadriven project from first the first idea to the deployed prototype.

### Goals

Two measure the success of this project certain goals and requirements were imposed on the projectteam. Some of these goals and requirements were set by Dr. Anderas Reschreiter as part of the assignment other goals the project team set for them self.

The requirements for the assignment were to develop a model that can predict a taxi fare based on the input of two addresses (pickup and drop off) as well as a time when the taxi should pick up a potential customer. Therefor a shiny app should be developed that allows a user to input these information quickly.

The goals the team set for them self are described below.

### Qualitative Objectives

- Develop a predictive model to estimate the fare amount for taxi trips in Chicago
- Evaluate various modeling approaches and deploy the best solution

### Quantitative Objectives

- Achieve **RM (SE per 10 minutes) < $2** on the validation set
- Achieve **mean AE/fare < 5%** for on the validation set
- Compare performance across 5 different model types

### Models

The models that were considered and analysed in this report were chosen for their effectiveness with large datasets. All models were tested and compared with the same quantitative objectives.

- Neural network
- XGBoost

- Random Forest
- Linear Model
- GLM

## Planned Process

The process of this project and report can be separated in 6 steps.

1. data aquisition
2. data cleaning
3. dataset preparation
4. model training and evaluation
5. ui development
6. combining the final works

Each of these steps is discussed in a separate chapter. During the development and project phase multiple of those steps were done simultaneously to enable a swift and fast process progression.

And estimated project timeline can be seen below

[ insert timeline ]

## Data acquisition

### source

### raw data

### challenges

## Data cleaning

### data completeness

### error and impossible measurements

## Dataset preparation

### column selection

### stratifying process

### dataset sizes

### Model training and evaluation