Machine Learning in Finance (Seminar)
Group Project Outline

Benjamin Zimmermann

March 2023

1 Task Description

1.1 Goal

Any application in machine learning (ML) is ultimately a hands-on task. Therefore the underlying objective of this group project is that students learn to apply the discussed ML algorithms and techniques using Python as programming language. To accomplish this, groups are to use a data set with explanatory variables to derive real estate price (classes).

1.2 Data Source/Set

The data for this group project is taken from OpenML's free datasets and can be accessed through OLAT. The following link provides further description of the data: `https://www.openml.org/search?type=data&sort=runs&id=42165&status=active`

Note that the original response variable is numeric and used in regression task settings. However, for the purpose of this seminar the response variable `SalePrice` was transferred into price-range bins/classes called `Class`.

1.3 Task: Real Estate Price Class Estimation

The data set provides 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa. Students are to use the given data to train a model that is able to produce price bin/class estimations. The response classes[1] are:

- **0**: Sale Price between \$0 and \$100,000
- **1**: Sale Price larger than \$100,000 but smaller than or equal to \$200,000
- **2**: Sale Price larger than \$200,000 but smaller than or equal to \$300,000
- **3**: Sale Price larger than \$300,000 but smaller than or equal to \$400,000
- **4**: Sale Price larger than \$400,000

This tasks implies dealing with the following questions:

- How does one handle missing data and outliers?
- Selection of relevant explanatory variables (features) for your model: Does a model that includes all available 79 features yield the best results or is a restricted model with only a subset more convincing?

[1] For simplicity reasons the task is limited to 5 classes; obviously any real application would have to be more granular (or alternatively a regression task).

- Feature engineering: Does adding other or creating new features (besides the given fundamental data) improve your model (e.g. macroeconomic or demographic data etc.)?
- Class imbalance: Is class imbalance affecting accuracy?

2 Grading

2.1 Requirements

The case study requirements are:

1. **Code/script (Jupyter notebook) & results** (50% of grade)
   - Reproducibility: Can code be run on lecturer's computer without problem?
   - Comprehensibility: Is code well documented, can one follow the thought process?
   - Adequacy: Is (thought) process in-line with econ/finance theory and is the reasoning well documented? Adequacy of methods to handle missing/erroneous data?
   - Performance: How precise is the group's recommendation model (chose relevant measures such as accuracy, confusion matrix, F_1-Score, other relevant performance metrics to rate your own results)?
2. **Abstract/Summary** (3-5 pages; 25% of grade)
   Here you should briefly outline your final setup (algorithms, features selection etc.), how you dealt with `NaN` values, outliers, and class imbalance, your results, how convinced you are of your results, what further analysis you see necessary, and any literature you considered.
3. **Presentation of results and Q&A** ($\leq$ 10 slides, exactly 10 min; 25% of grade)
   Present your results (similar to what you discuss in Abstract/Summary) and be prepared to discuss these and your code during a Q&A.

All documents (scripts, presentation, report, data, other relevant files) are to be handed in no later than 16.04.2023 through a cloud repository ($\rightarrow$ **time stamp must be uncompromisable**). Language: either English or German, the latter is only allowed if all group members agree on it.

One grade will be assigned per group and all group member need to be present during the presentation.

2.2 Presentation Schedule

All group presentations are in-person at the University of Zurich (room KOL-F-103) on Tue, 25.04.2023. The schedule is as follows:

Start Time	Group	Members
16:15	1	
16:45	2	
17:15	3	See OLAT and/or attached overview
17:45	4	
18:15	5	

Universität Zürich[UZH]

Groups "Introduction to Machine Learning"

Surname	First name	Email	Group
Beck	Alexander	*alexander.beck2@uzh.ch*	1
Blank	Joel	*joel.blank@uzh.ch*	1
Müntener	Pascal	*pascal.muentener@uzh.ch*	1
Öztürk	Kenan	*kenan.oeztuerk@uzh.ch*	1
Pavlics	Arthur	*arthur.pavlics@uzh.ch*	1
Gong	Yiwei	*yiwei.gong@uzh.ch*	2
Greb	Cléo	*cleo.greb@uzh.ch*	2
Kägi	Pascal	*pascal.kaegi2@uzh.ch*	2
Malovecky	Juraj	*juraj.malovecky@uzh.ch*	2
Niklaus	Morris	*morris.niklaus@uzh.ch*	2
Barandun	Benedict	*benedict.barandun@uzh.ch*	3
Maznichenko	Lev	*lev.maznichenko@uzh.ch*	3
Simtion	Victor Nicolae	*victornicolae.simtion@uzh.ch*	3
Tran	Minh Hien	*minhhien.tran@uzh.ch*	3
Walder	Anne-Catherine Sophia	*anne-catherinesophia.walder@uzh.ch*	3
Herman	Janice	*janice.herman@uzh.ch*	4
Leuthard	Valentin	*valentin.leuthard@uzh.ch*	4
Reding	Andrin	*andrin.reding@uzh.ch*	4
Reding	Nico	*nico.reding@uzh.ch*	4
Tessendorf	Liam	*liam.tessendorf@uzh.ch*	4
Kuljanin	Aleksandar	*aleksandar.kuljanin@uzh.ch*	5
Schlegel	Arthur	*arthur.schlegel@uzh.ch*	5
Stoll	Cyrill	*cyrill.stoll@uzh.ch*	5
Waber	Selina Svenja	*selinasvenja.waber@uzh.ch*	5