



Universität St.Gallen

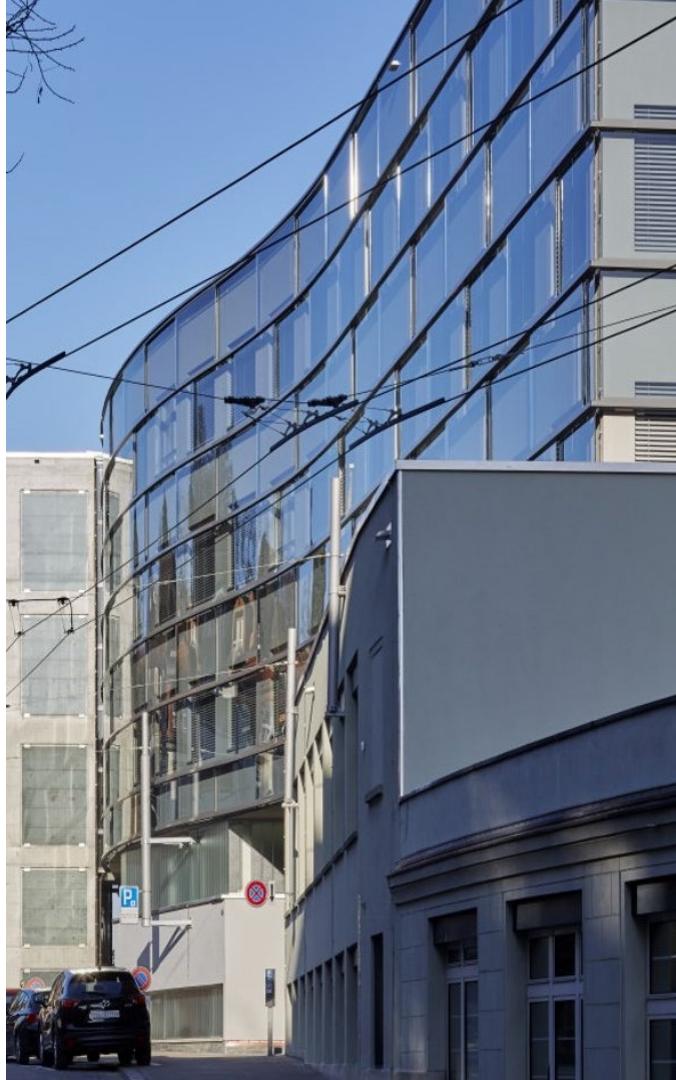
Institut für Wirtschaftsinformatik

Submitting on Kaggle Competitions

Prof. Dr. Ivo Blohm (ivo.blohm@unisg.ch)

Associate Professor for Information
Systems and Business Analytics

From insight to impact.



Competition Website

Aufgabe 3 - Gebrauchtwagenpreise

Verpflichtend: Assignment 3 & Teilnahme an Kaggle Wettbewerb

Der Gebrauchtwagenmarkt befindet sich gerade in einer Phase der digitalen Transformation. Craigslist ist einer der führenden Anbieter von Gebrauchtwagenanzeigen und möchte nun die eigenen Dienstleistungen in diesem Bereich ausbauen. Sie wurden engagiert ein entsprechend prädiktives Modell zu entwickeln, mit denen Gebrauchtwagenpreise vorhergesagt werden können, um Käufern und Verkäufern den Transaktionsprozess zu erleichtern.

- Details zur Aufgabe: [Assignment3.pdf](#) ↴
- Datensatz: [assignment.csv](#) ↴

Für das Assingment sind die Self-Studies zu Modul 4 und Modul 5 relevant.

Wie gut sind die eigenen Vorhersagen im Vergleich zu denen der Mitstudierenden? Finden Sie es heraus. Es gibt einen zweiten Datei (kaggle csv), für den die Preise nicht verfügbar sind, d.h., wir haben ausschliesslich die Features zum Erstellen der Vorhersagen - wir aber nicht evaluieren. Die Preise für die Autos im kaggledata-Dataset sind auf der [Kaggle Competition Website](#) hinterlegt und mit dem Notebook können auf Basis Ihrer im Assignment 3 erstellen Modelle Prognosen erstellt und ein formatiertes "Submission file" erstellt werden. Kaggle hochgeladenen Vorhersagen werden auf Basis des RMSE evaluiert und anschliessend in einem Leaderboard dargestellt.

Der Kaggle-Wettbewerb ist **nicht öffentlich** und nur über diesen Link zu erreichen. Teilnehmer sind ausschliesslich aus dem Kurs 8.01f 2023. Für die Teilnahme ist ein google account oder ein kaggle account notwendig. [Die Teilnahme am Wettbewerb hat keine Auswirkungen auf die Note](#).

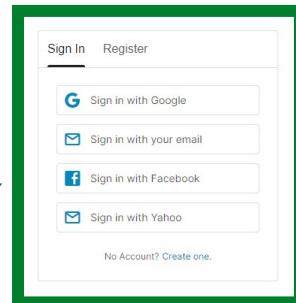
Bewertung des Assignment 3:

- Notebook zum Erstellen formatierter Vorhersagen: [CreateKaggleSubmission.ipynb](#) ↴
- Datensatz für Wettbewerb: [kaggle.csv](#) ↴
- Beschreibung Einreicheprozess: [kaggle.pdf](#) ↴
- [Kaggle Wettbewerb](#) ↴

Freiwillig: Abgabe mit Echtnamen

Veröffentlichen

kaggle



Create Login / Login
with Google Account
Or Email

Competition Website

The screenshot shows a web browser displaying a competition page on Kaggle. The URL in the address bar is kaggle.com/competitions/mbl2022Cup. The left sidebar contains navigation links such as Create, Home, Competitions, Datasets, Code, Discussions, Courses, and More. Under Your Work, recently viewed competitions include MBI Big Data & Data Science Competition, CAS BDAI Cup 2020, CAS BDAI Cup, and Heart Failure Prediction. A search bar at the top right has the placeholder "Search". The main content area features a banner for the "Community Prediction Competition" titled "MBI Big Data & Data Science Competition" with the sub-question "Who can best predict used car's prices?". It shows a photo of several cars and people. Below the banner, a green button labeled "Join Competition" is highlighted with a green box. A horizontal menu bar below the banner includes Overview, Data, Code, Discussion, Leaderboard, and Rules. The Overview section contains a "Description" box stating: "Craigslist is the world's largest collection of used vehicles for sale, yet it's very difficult to collect all of them in the same place. I built a scraper for a school project and expanded upon it later to create this dataset which includes every used vehicle entry within the United States on Craigslist." It also contains an "Evaluation" box with the instruction: "Can you predict the price of the used cars listed on craigslist? Build a model that minimizes RMSE!". At the bottom of the main content area, there is a timeline showing "Launch 4 minutes ago" and "Close a month".

Join Competition

Verifiy your account

The screenshot shows a web browser window with a Kaggle competition page in the background. Two modals are overlaid on the page:

- Modal 1: Just one thing—first verify your account**

Just one thing—first verify your account

Enter your phone number and we'll send you a code

COUNTRY: CH +41 CH | PHONE NUMBER: Phone number

I'm not a robot

Verifying with a phone number helps us prevent spam and fraud on Kaggle.

 - You can only have one Kaggle account. If you have another account, you'll need to first delete that account.
 - The phone number needs to be yours, and not a public or shared number.
 - Message and data rates apply.

Contact us for help >

Cancel Send verification code

4 minutes ago 8 months
- Modal 2: Please read the competition rules**

Please read the competition rules

CAS BDAI Cup Rules

By clicking on the "I Understand and Accept" button below, you agree to be bound by the competition rules.

I Understand and Accept

Accept Competition Rules

Make Submissions

The screenshot shows the Kaggle interface for the 'MBI Big Data & Data Science Competition'. The main banner features a photo of people at a car show with the text 'Community Prediction Competition' and 'MBI Big Data & Data Science Competition'. Below the banner, a 'Submit Predictions' button is highlighted with a green box. The navigation bar includes links for Overview, Data, Code, Discussion, Leaderboard, Rules, Team, Host, My Submissions, and an ellipsis. The left sidebar lists various sections like Home, Competitions, Datasets, Code, Discussions, Courses, and More. Under 'Your Work', it shows recently viewed competitions: MBI Big Data & Data Science Competition, CAS BDAI Cup, Car Prices Dataset, Used Cars Price Pre..., How to: Folium for ..., and View Active Events.

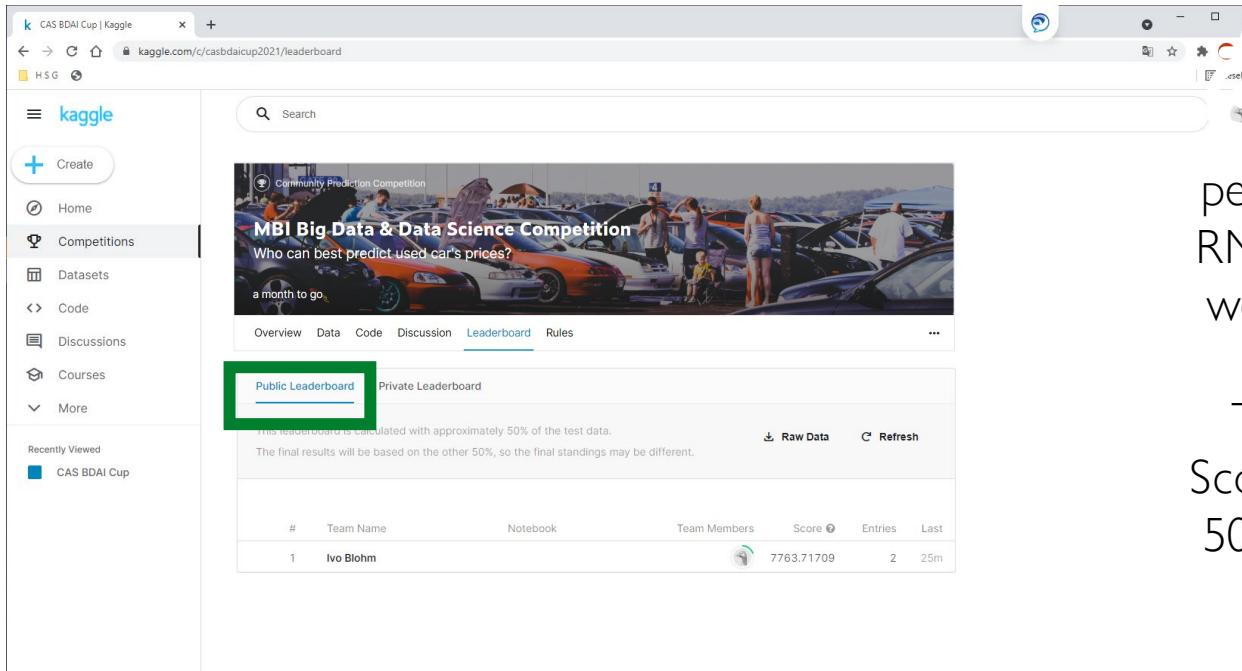
Take the CAS BDAI CUP
Notebook

Develop your Model in
Google Collab / Jupyter
Notebook

Create Solution File
and submit it here

10 submissions per
day are allowed

Compare your Performance – Public Leaderboard



The screenshot shows a web browser window for the Kaggle platform, specifically the 'CAS BDAI Cup' competition. The left sidebar contains navigation links like 'Create', 'Home', 'Competitions', 'Datasets', 'Code', 'Discussions', 'Courses', and 'More'. The main content area displays a banner for the 'MBI Big Data & Data Science Competition' with the tagline 'Who can best predict used car's prices?'. Below the banner, tabs for 'Overview', 'Data', 'Code', 'Discussion', 'Leaderboard' (which is underlined in blue), and 'Rules' are visible. The 'Leaderboard' section is titled 'Public Leaderboard' and shows a table with one entry:

#	Team Name	Notebook	Team Members	Score	Entries	Last
1	Ivo Blohm			7763.71709	2	25m

Below the table, a note states: 'This leaderboard is calculated with approximately 50% of the test data. The final results will be based on the other 50%, so the final standings may be different.' There are also 'Raw Data' and 'Refresh' buttons.

After submitting a contribution, the performance get scored by RMSE and you can see how well your solution is doing

The Public Leaderboard Score is calculated based on 50% of the «ground truth» data

Compare your Performance – Private Leaderboard

CAS BDAI Cup | Kaggle

MBI Big Data & Data Science Competition

Who can best predict used car's prices?

a month to go

Overview Data Code Discussion Leaderboard Rules

Public Leaderboard Private Leaderboard

This leaderboard is calculated with approximately 50% of the test data. The final results will be based on the other 50%, so the final standings may be different.

Raw Data Refresh

#	Team Name	Notebook	Team Members	Score	Entries	Last
1	Ivo Blohm			7763.71709	2	25m

Performance is scored on the other half of the «ground truth» data

Performance scores are hidden (also for submitters)

Private Leaderboard determines winner

Why so complicated? Target Leakage!

Presence of **unexpected additional information** in the training data, allowing a model to make **unrealistically good predictions**.

- Features in historic data that would not be available at the time of the prediction
- Duplicated instances of test data in training data
- Predictions get part of test data

Model looks good in development, but performs poor in the real-world

If contestants know their performance on the „evaluation data“ they would tailor their models to this specific data set!

HAPPY

HACKING