



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

Veröffentlichen

kaggle

### Verpflichtend: Assignment 3 & Teilnahme an Kaggle Wettbewerb

Der Gebrauchtwagenmarkt befindet sich gerade in einer Phase der digitalen Transformation. Craigslist ist einer der führenden Anbieter Gebrauchtwagenanzeigen und möchte nun die eigenen Dienstleistungen in diesem Bereich ausbauen. Sie wurden engagiert ein entsprechendes 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](#) ↓

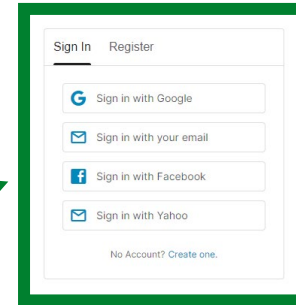
Für das Assignment sind die Self-Studies zu Modul 4 und Modul 5 relevant.

Wie gut sind die eigenen Vorhersagen im Vergleich zu denen der Mitstudierenden? Finden Sie es heraus. Es gibt einen zweiten Datensatz (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 erstellten Modelle Prognosen erstellt und ein formatiertes "Submission file" erstellt. 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 8,014 2023. Für die Teilnahme ist ein Google Account oder ein Kaggle Account notwendig. **Die Teilnahme am Wettbewerb hat keine Auswirkung auf die Bewertung des Assignment 3.**

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

Freiwillig: Abgabe mit Echtnamen

A screenshot of the Kaggle login/register page. It features a green border around the login area. At the top, there are tabs for 'Sign In' and 'Register'. Below these are four buttons: 'Sign in with Google' (with a Google logo), 'Sign in with your email' (with an email icon), 'Sign in with Facebook' (with a Facebook logo), and 'Sign in with Yahoo' (with a Yahoo logo). At the bottom, there is a link that says 'No Account? Create one.'

Create Login / Login  
with Google Account  
Or Email

# Competition Website

The screenshot shows the Kaggle website interface. On the left is a sidebar with navigation links: Home, Competitions, Datasets, Code, Discussions, Courses, More, Your Work, and View Active Events. The main content area features a search bar and a large banner for the 'MBI Big Data & Data Science Competition'. The banner includes the text 'Who can best predict used car's prices?' and 'a month to go'. Below the banner are tabs for Overview, Data, Code, Discussion, Leaderboard, and Rules. A 'Join Competition' button is highlighted with a green box. The Overview section contains a description of the competition, an evaluation section, and a timeline showing the launch and close dates.

Join Competition

# Verify your account

MBI Big Data & Data Science Co

kaggle.com/competitions/mbi2022Cup

**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

Already have a code? [Enter it now.](#)

[Cancel](#) [Send verification code](#)

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 >](#)

[Terms and Conditions](#) · [Privacy](#)

**Please read the competition rules**

[CAS BDAI Cup Rules](#)

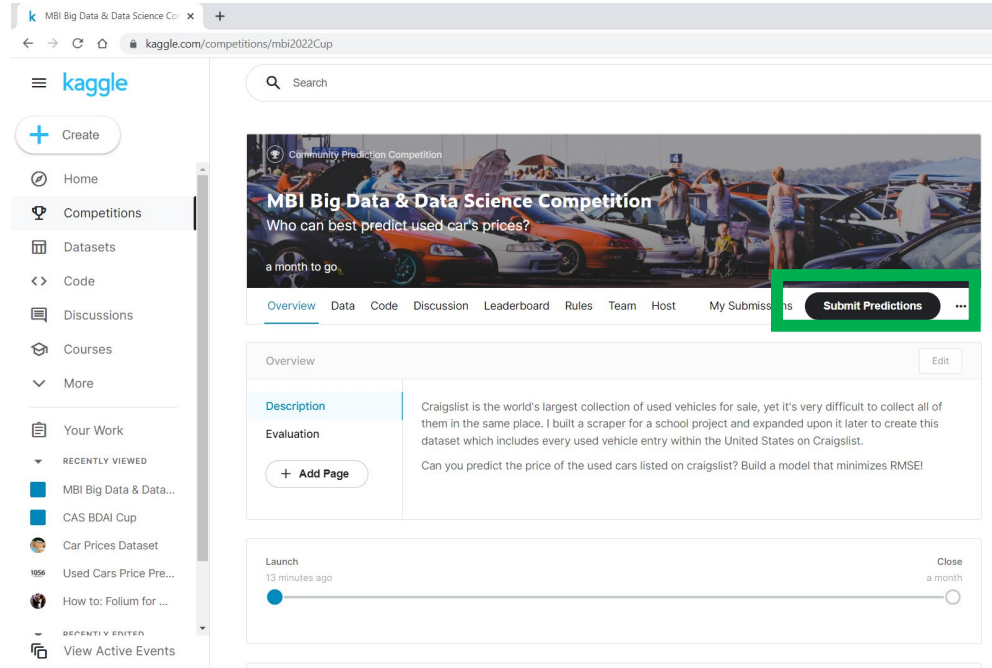
By clicking on the "I Understand and Accept" button below, you agree to be bound by the competition rules.

[I Understand and Accept](#)

2 Points This competition does not award ranking points

## Accept Competition Rules

# Make Submissions



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 the Kaggle website interface. The left sidebar contains navigation links: Home, Competitions, Datasets, Code, Discussions, Courses, and More. The main content area displays the 'MBI Big Data & Data Science Competition' with a banner image of people at a car show. Below the banner, the 'Leaderboard' tab is selected. A green box highlights the 'Public Leaderboard' link. The leaderboard table shows the top performer, Ivo Blohm, with a score of 7763.71709. The table has columns for Rank, Team Name, Notebook, Team Members, Score, Entries, and Last Update.

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

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

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

Performance scores are hidden (also for submitters)

Private Leaderboard determines winner

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

# 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