CV Of Tobias Klein

Email: progress.unveiled@gmail.com
Portfolio Website: https://deep-learning-mastery.com
GitHub Profile: https://github.com/kletobias

Contents

Currently

In my current position, I participate in machine learning competitions and expand my toolkit on a full-time basis.

Specialized In Machine Learning

To solve machine learning problems, I use Python and an iterative approach. In general, that involves the following steps, possibly going back and forth between them. For more information, see my article: **The Tasks In Every Machine Learning Project: Tabular Data**.

- Define the problem
    - Is it a regression problem
    - Is it a classification problem
    - Is it other higher order problem type
- Prepare Data
    - Data Preprocessing
    - Feature Selection
    - Feature Engineering
- Models
    - Candidate Model Selection
    - Hyperparameter Optimization (optional: depending on model)
    - Model Evaluation/Interpretation
    - Finalize Model

Toolkit

Programming Languages

Name	Experience	Examples
Bash	Advanced	Manage a Linux server from the command line using ssh, e.g., GPU cloud computing
C++	Basic	Initialize variable, loops, 'cout', user input, vectors
MySQL	Intermediate	Writing standard queries, creating and altering tables
Python	Expert	Entire ML workflow utilizing custom functions, reproducible code, shell scripts
R	Basic	Tidy, ggplot2, Multivariate Regression

Python Toolkit

Name	Type	Description	Proficiency
conda/pip/pyenv/virtualenv	Tools	Package and Python environment management	7/10
fastai	Library	Deep Learning Library based on PyTorch	7/10
numpy	Library	Data Manipulation using Vectorization	6/10
pandas	Library	Tabular Data Manipulation	8/10
pytorch	Library	Deep Learning Framework	5/10
pyplot	Library	General Purpose Data Visualization	8/10
pytorch_tabular	Library	Tabular Data Deep Learning based on PyTorch	4/10
re	Library	Builtin regular-expressions library	8/10
scikit-learn	Library	Multipurpose ML Library	7/10
scipy	Library	Mainly used the scipy.stats module	4/10
seaborn	Library	Statistical Data Visualization	7/10
tpot	Library	Automated Machine Learning tool	6/10
xgboost (dmlc)	Library	Regression and Classification ML model	8/10

Language Proficiency

- German: Native
- English: Bilingual Proficiency
    - The most recent English proficiency test (Toefl IBT) was taken in 2019 with a score of 111/120.
- Spanish: Elementary Proficiency

School

`1993 – 1994` **Lincoln Elementary School**, Kampala, Uganda

`1994 – 1995` **Kindergarten**, Freiburg i. Br.

`1995 – 1999` **Elementary School**, Zell a.H.

`1999 – 2006` **High School**, Freiburg i. Br.

`2006 – 2006` **Secondary School**, Whistler, Canada

- Full-time mountain bike junior development program with participation in races.

`2006 – 2007` **High School**, Freiburg i. Br.

`2007 – 2008` **High School**, Hamburg

- Graduated with Abitur.

Higher Education

`2009 – 2011` **Heinrich Heine University**, Düsseldorf

- **Subject:** Law (Jura)
- **Optional Information:**
  - **Courses Completed**
    - Zivilrecht: BGB AT, Schuldrecht AT & BT, Hausarbeit
    - Strafrecht: STGB AT I-II, Hausarbeit im Strafrecht, Übung im Strafrecht
    - Öffentliches Recht: Polizeirecht, Grundrechte, Allg. Verwaltungsrecht, Verwaltungsprozessrecht
    - 1. Teil Kurs im Angloamerikanischen Recht

`2011 – 2016` **LMU/Freiburg University**, Munich/Freiburg i. Br.

- **Subject:** Mathematics B.Sc.
- **Optional Information:**
  - **Courses Completed**
    - Linear Algebra I-II | Grades: {'I': 1.7, 'II': 2.3}
    - Analysis I-III | Grades: {'I': 1.3, 'III': 1.7}
    - Stochastic | Grade: 4.0
    - Complex Analysis | Grade: 4.0
    - Futures And Options | Grade: 3.7

    - Exercise In Numerics Using **C** To Implement Methods From Linear Algebra.

`2016 – 2019` **Freiburg University**, Freiburg i. Br.

- **Subject:** Economics Focused Business Administration (BWL Non-Profit & Public Management B.Sc.)
- **Final Grade:** 1.6
- **Bachelor Thesis:**
    - **Title:** *'Data Mining: Hyperparameter Optimization For Real Estate Prediction Models'*
    - **Written Using:** *Latex* & *Python*
    - **Pages:** 69
    - **Grade:** 1.0
    - **Abstract:** See Section 'Bachelor Thesis'
    - **Full Text (pdf)**

- **Optional Information:**
    - **Table: Relevant Courses Completed & Grade**

Course	Grade
Bachelor Thesis 'Data Mining: Hyperparameter Optimization for Real Estate Prediction Models'	1.0
Business Intelligence	1.3
Econometrics	1.0
Electives in Non-Profit Management	2.3
Foundations of Economic Policy	1.0
Fundamentals of Public Management: Foundation of Public Management	1.3
Game Theory: Spieltheorie	1.0
Health Care Management I: Gesundheitsmanagement I	2.3
Human Resources and Organization: Human Resources and Organisation	2.3
Introduction to Information Systems	1.0
Law & Economics	2.0
Management and Theory of the Firm: Unternehmenstheorie	1.3
Mathematics	1.3
Microeconomics I	1.3
Microeconomics II: Mikroökonomik II	1.7
New Public Management	1.3
Non-Profit Organizations	1.3
Public Finance I	2.0
Public Finance II	1.0
Tax Management (Seminar)	1.3

Bachelor Thesis

Data Mining: Hyperparameter Optimization For Real Estate Prediction Models

Abstract:
Combining a highly scalable and customisable process, with very accurate prediction results using machine learning models, is what this work proposes. The customisation is guided by what information the user seeks to gain from the process. This makes the process applicable for a variety of sectors, such as Banking & Finance, Marketing and urban development among others. It evaluates the process of using self-acquired

data from an online real estate platform, gained from deploying a custom web scraping algorithm. This data is then combined with several spatial features for predicting the base rent for apartments on a validation dataset. The analysis and predictions are made for rental apartment listings within the Hanseatic City of Hamburg. The spatial features originate from sources other than that of the apartments data and have to be adapted to it first, therefore. Predictions are made using state of the art machine learning models, in the form of a **Lasso Regression** model and a **XGBoost Regressor** model. The Hyperparameter Optimisation techniques **grid search** and **random search** are compared, during the optimisation process. The focus is on maximising prediction accuracy of the models. The best scores, expressed in **RMSE**, are 190.68 for the **Lasso** and 115.39 for the **XGBoost Regressor**. Differences in complexity and interpretability between the models are discussed and associated with it, the strengths and weaknesses of the respective model are pointed out.
Full Text (pdf)

Portfolio Articles

Listed below are the portfolio articles published on my website, accompanied by descriptions and tags, which provide a concise description of their content in terms of relevant search terms. Additionally, the article's category, word count, and link to the full text are included.

Automation Using A Test Harness For Deep Learning: Part 1

Description: How to create and use a custom test harness, that automates many steps of the deep learning testing process. It lowers GPU idle time, lets one build more models, test more parameter combinations in less time. The fastai library for deep learning is used throughout this article.
Tags: ['binary-classification', 'deep-learning', 'fastai', 'hyperparameter-optimization', 'learning-rate', 'loss-function', 'stochastic-gradient-descent']
Category: *deep-learning* | **Word Count:** 2703 | **Full Article**

Automation Using A Test Harness For Deep Learning: Part 2

Description: This is Part 2 in the series, where we explore how the fastai deep learning library can be used to conduct structured empirical experiments on a novel and small dataset. The dataset consists of 850 images and an almost uniform distribution for the target labels. There are two labels in total, "male" and "female", that are assigned the gender of the model depicted in any of the images in the dataset.
Tags: ['binary-classification', 'deep-learning', 'fastai', 'hyperparameter-optimization','image-data']
Category: *deep-learning* | **Word Count:** 3060 | **Full Article**

Advanced Geospatial Feature Creation

Description: Extensive cleaning and transformation of tabular data, in order to create geospatial features. Once processed, the results are clean GPS values as "Point" objects in decimal degrees format and names of all subway and suburban train stations within Hamburg, Germany.
Tags: ['data-cleaning', 'data-transformation', 'geospatial-feature-creation', 'regular-expression', 'shapely', 'tabular-data']
Category: *data-preprocessing* | **Word Count:** 3777 | **Full Article**

Cleaning a web scraped 47 Column Pandas DataFrame Part 1

Description: Data Preparation Series: Exploring Tabular Data With pandas: An Overview Of Available Tools In The pandas Library.
Tags: ['data-exploration', 'first-steps', 'introduction', 'pandas', 'tabular-data']
Category: *data-preprocessing* | **Word Count:** 3444 | **Full Article**

Cleaning a web scraped 47 Column Pandas DataFrame Part 2

Description: More efficient string data cleaning by using the pyjanitor module and method chaining.
Tags: ['data-cleaning', 'pandas', 'regular-expressions', 'string-manipulation', 'tabular-data']
Category: *data-preprocessing* | **Word Count:** 3199 | **Full Article**

Cleaning a web scraped 47 Column Pandas DataFrame Part 3

Description: Extensive cleaning and validation and creation of a valid GPS column from the records, by joining the longitude and latitude columns together using geometry object Point.
Tags: ['data-validation', 'dtype-timedelta64','geospatial-feature-engineering', 'pandas', 'tabular-data']
Category: *data-preprocessing* | **Word Count:** 2339 | **Full Article**

Cleaning a web scraped 47 Column Pandas DataFrame Part 4

Description: Extensive data cleaning and validation using regular expressions. Showcase of how batch processing several columns of tabular data using pandas, pyjanitor and the re library can look like.
Tags: ['batch-processing', 'data-validation', 'pandas', 'regular-expressions', 'tabular-data']
Category: *data-preprocessing* | **Word Count:** 4193 | **Full Article**

MySQL Queries Using An AWS Redshift MySQL Database

Description: This article shows how one can use Python to import CSV files using Pandas into a MySQL database hosted on AWS using Redshift and how to formulate basic MySQL queries to get the data of interest.
Tags: ['mysql', 'AWS', 'pandas', 'tabular-data', 'query']
Category: *tabular-data* | **Word Count:** 1982 | **Full Article**

Datacamp Concrete Regression Challenge

Description: This is the notebook I created to solve the datacamp concrete challenge within an hour. There are explanations for most of the code in this article and we look deeper into the workings of the Lasso regression model.
Tags: ['cross-validation', 'lasso-regression', 'math', 'multivariate-regression', 'regression-analysis']
Category: *tabular-data* | **Word Count:** 1162 | **Full Article**

The Tasks In Every Machine Learning Project: Tabular Data

Description: The six tasks in every machine learning project with structured data.
Tags: ['predictive-modeling', 'hyperparameter-optimization', 'reproducable-code', 'tabular-data', 'feature-engineering']
Category: *tabular-data* | **Word Count:** 579 | **Full Article**

Deep Dive Tabular Data Pt. 1

Description: Preprocessing Data: Visualizing missing values on an 80 feature dataset. Strategies for filling missing values and using categorical embeddings.
Tags: ['categorical-embeddings', 'data-preprocessing', 'fastai', 'fill-strategies', 'tabular-data']
Category: *tabular-data* | **Word Count:** 5305 | **Full Article**

Deep Dive Tabular Data Pt. 2

Description: Feature selection using model DecisionTreeRegressor from sklearn and the feature_importances_ method which is tested for deviations in its score.
Tags: ['decision-tree-regressor', 'feature-importance', 'feature-selection', 'sklearn', 'tabular-data']
Category: *tabular-data* | **Word Count:** 1839 | **Full Article**

Deep Dive Tabular Data Pt. 3

Description: RandomForestRegressor using feature_importances_ and out-of-bag error to asses model performance.
Tags: ['feature-importance', 'feature-selection', 'out-of-bag-error', 'random-forest', 'tabular-data']
Category: *tabular-data* | **Word Count:** 1351 | **Full Article**

Deep Dive Tabular Data Pt. 4

Description: Interpretation Using Advanced Statistical Visualizations. Dendrogram, Spearman rank correlation, partial dependence plot, impact of independent variables for sample on predictions.
Tags: ['dendrogram', 'partial-dependence', 'spearman-rank-correlation', 'tabular-data', 'treeinterpreter']
Category: *tabular-data* | **Word Count:** 1607 | **Full Article**

Deep Dive Tabular Data Pt. 5

Description: Out-of-domain problem: What it is, why it is important, how to spot it and how to deal with it.
Tags: ['feature-importance', 'model-accuracy', 'out-of-domain-problem', 'random-forest', 'tabular-data']
Category: *tabular-data* | **Word Count:** 1030 | **Full Article**

Deep Dive Tabular Data Pt. 6

Description: Kaggle Submission 1: Training RandomForestRegressor, fastai deep learning model using hyperparameter optimization techniques. Preprocessing of Kaggle test data.
Tags: ['data-preprocessing', 'fastai', 'hyperparameter-optimization', 'random-forest', 'tabular-data']
Category: *tabular-data* | **Word Count:** 2647 | **Full Article**

Deep Dive Tabular Data Pt. 7

Description: Kaggle Submission 2: tabular_learner deep learning estimator optimized using manual hyperparameter optimization. XGBRegressor using RandomizedSearchCV and sampling from continuous parameter distributions.
Tags: ['hyperparameter-optimization', 'random-search', 'tabular-data', 'tabular_learner', 'xgboost-regressor']
Category: *tabular-data* | **Word Count:** 3459 | **Full Article**

The Math Behind "Stepping The Weights"

Description: In this article we highlight a key concept in the Stochastic Gradient Descent and explore the basics, that this optimization algorithm is derived of.
Tags: ['deep-learning', 'math', 'ordinary-least-squares', 'partial-derivate', 'stochastic-gradient-descent']
Category: *deep-learning* | **Word Count:** 1756 | **Full Article**