**Masterarbeitstitel**

Vergleich der Vorhersagegüte von neuronalen Netzen und baumbasierten Algorithmen bei Multiklassifikation basierend auf Worteinbettungen der Verwendungszwecke von finanziellen Transaktiondaten.

Comparison of prediction accuracy between neural nets and tree-based algorithms based on wordembeddings of purpose titles of financial transaction data

**Preprocessing:**

* **Bag of words**
* **Andere**
* **dimensionsreduktion**

**Datensatz beschreibung:**

* **In Raw**
* **Textmining preprocessing**

**Exploration:**

* **Mengengerüst der klassen**
* **Häufigste Wörter pro Klasse**

**Features:**

* **Number of Camel words**
* **Number of camel words**
* **Number of total words**
* **Bag of words**

**Train/Test Split**

* **Kategorien gebalanced im Train Set**
* **Nachschauen über binomialverteilung wieviele Klasse im Testset nötig sind**

**Methoden**

* **Multinomial logit mit /ohne spline: als eta alles mögliche, zB auch predictor vom neuronalenNetz**
* **Bamlss paket: erweiterungen**
* **Classification RandomForest**
* **XGBoost**

**Performance maß:**

* **rank probability score,**
* **difference between multinomial likelihood and true classes**
* **brier-score**

**Next Steps:**

* **Buch zu categorial data holen und multinomial logit lesen**
* **Bamss Paper lesen**
* **Datensatz einlesen und Framework schreiben**
* **David Middelbeck empfehlung:** Sandbox transaktionsdaten schauen
* **Performance Maße nachschauen**
* **Bei Kaggle schauen was bisher auf dem Datensatz gemacht wurde**
* **LSTM, HAN lessen, generell text classification**

tags:

textmining word embeddings multicategorial transaction data neural nets comparison xgboost

.

daten müssen enthalten:

transaktionsdaten:

* Verwendungszweck, auftraggeber, betrag
* empfänger: zb amazon marketplace
* Zielvariable: categorie zb. online einkauf güter (40 Klassen +)
* Zielvariable: unterkategorie: elektronik