Present the state of the art using the literature review

Time series analogies (REMA) vs the newer neural networks

Competition, Github – sources

Unlicensed data is not allowed.

Time series are good for live stock trading. The trained model needs to be retrained every 30min otherwise the loss becomes too high.

Another example for this is the weather forecast.

I had a look at x.y research paper and Ive tried this however I think that this is the case

Set the referencing style to Harvard in word.

Independent features label or mapped to dependent features

Adam is the best optimizer (indicator that the optimizer is not working is that the loss function is not reducing)

<https://www.met.ie/climate/available-data/long-term-data-sets>

<https://www.fiscalcouncil.ie/datasets/>

<https://data.gov.ie/dataset?q=PRICE&theme=Housing&sort=score+desc%2C+metadata_created+desc>

kaggle.com/code/burhanykiyakoglu/predicting-house-prices/notebook

<https://www.kaggle.com/datasets/sudalairajkumar/cryptocurrencypricehistory/discussion>

<https://www.kaggle.com/datasets/nancyalaswad90/diamonds-prices>

<https://www.academia.edu/71572423/Knowledge_Graphs_and_Big_Data_Processing?from_sitemaps=true&version=2>

# <https://www.sciencedirect.com/science/article/abs/pii/S0167923610001430> (An intraday market risk management approach based on textual analysis)

The future will most probably belong to those companies that can turn data processing into new technologies and products. Application for machine learning using deep learning and bid data are limitless. A new protein that stops or slows ageing process can be discovered or a new type of propulsion motor using a new type of fuel or automate most of the jobs.

Big data can basically be classified into two categories, namely, data from the physical world, which is usually obtained through sensors, scientific experiments and observations (such as biological data, neural data, astronomical data, and remote sensing data), and data from the human society, which is often acquired from such sources or domains as social networks, Internet, health, finance, economics, and transportation.

Actually, the real challenges center around the diversified data types (Variety), timely response requirements (Velocity), and uncertainties in the data (Veracity).

semi-structured or unstructured data (including text, images, video, and voice)

Timely responses are also challenging because there may not be enough resources to collect, store, and process the big data within a reasonable amount of time. Finally, distinguishing between true and false or reliable and unreliable data is especially challenging, even for the best data cleaning methods to eliminate some inherent unpredictability of data.

From the socio-economic point of view, big data is the core connotation and critical support of the so-called second economy, a concept proposed by the American economist W.B. Arthur in 2011 [12], which refers to the economic activities running on processor, connectors, sensors, and executors. It is estimated that by 2030 the size of the second economy will approach that of the first economy (namely, the traditional physical economy).

new and rich information, these aforementioned big data have served different prediction tasks in all the domains of

<https://www.datacamp.com/tutorial/xgboost-in-python>

<https://www.kaggle.com/code/testaccount93/mypersonality/notebook>

In a typical real-world project, you would want to spend a lot more time exploring the dataset and visualizing its features. But since this data comes built-in to Seaborn, it is relatively clean.

The dataset has three categorical columns. Normally, you would encode them with ordinal or one-hot encoding, but XGBoost has the ability to internally deal with categoricals.

The way to enable this feature is to cast the categorical columns into Pandas category data type (by default, they are treated as text columns):

A possible workaround is splitting the data into three sets. The model trains on the first set, the second set is used for evaluation and hyperparameter tuning, and the third is the final one we test the model before production.

But when data is limited, splitting data into three sets will make the training set sparse, which hurts model performance.

The solution to all these problems is cross-validation. In cross-validation, we still have two sets: training and testing.