Project Plan Notes

Armon

‘Deep neural networks’ – very successful in many applications, e.g. ‘images, audio and text’ – well able to ‘encode raw data into meaningful representations in these domains.’

‘Real-world applications, the most common data type is tabular data’.

‘Comprising samples (rows) with the same set of features (columns).’

‘Tabular data is used in practical applications in many fields, including medicine, finance, manufacturing, climate science, and many other applications that are based on relational databases.’

‘Traditional machine learning methods’ e.g. gradient boosting ‘still dominated tabular data modelling’ + better performance compared to ‘deep learning’.

Issues w neural nets – ‘lack of locality’, ‘data sparsity (missing values)’, ‘mixed feature types (numerical, ordinal and categorical)’ + ‘lack of prior knowledge about the dataset structure (unlike with text or images)’.

Potential for deep learning to perform better than decision-tree based in certain contexts.

‘Gradient boosting is an algorithm in which new models are created from previous models’ residuals and then combined to make the final prediction’. – ‘When adding new models, it uses a gradient descent algorithm to minimise the loss.’ – ‘XGBoost is one of the most popular GBDT implementations.’

Used XGBoost + deep learning.

Hyperparameter optimisation easier + results usually better for XGBoost.

Ensemble deep learning + XGBoost best.

‘No one type of model is always better or worse than any other model.’

‘Deep learning not currently all we need for tabular data’.

‘XGBoost is one of the most popular’ gradient-boosting method – also references LightGBM + CatBoost.

Used hyperopt.

Performance of neural networks purported in papers – not as generalisable as XGBoost – mainly performs well on dataset used in paper.

McElfresh (2024)

Debate between ‘NN vs GBDT (gradient-boosted decision trees)’ ‘overemphasised’.

‘For a surprisingly high number of datasets, either the performance difference between GBDTs and NNs is negligible, or light hyperparameter tuning on a GBDT is more important than choosing between NNs and GBDTs.’

Finds that – ‘GBDTs are much better than NNs at handling skewed or heavy-tailed feature distributions and other forms of dataset irregularities.’

Tabular data no consensus between NNs and GBDTs compared to ‘computer vision and natural language understanding, in which NNs have far outpaced competing methods.’

For many applications trying out many different methods (NN or GBDT) is unnecessary – well tuned GBDT model is sufficient.

Some applications where NNs are better – however, on average GBDTs performed better over the 176 datasets studied.

GBDTs perform relatively better on datasets with heavy tails + larger datasets – NNs better w regular datasets + smaller datasets.

‘GBDTs iteratively build an ensemble of decision trees, with each new tree fitting the residual of the loss from the previous trees, using gradient descent to minimize the losses.’

Grinsztajn (2022)

Uninformative features a factor in hindering neural nets.

Also again irregular feature distributions harm neural nets.

GBDTs lower ‘computational cost’.

Gorishniy (2021)

ResNet vs FT-Transformer NN architecture.

Used Optuna optimisation.

GBDTs not good in cases of multi-class w many classes.

Borisov

Tabular datasets are heterogeneous not homogeneous (‘dense numerical + sparse categorical features’) – correlation between features less strong than in ‘image or speech data’ – causing issues with neural networks.

High flexibility – benefit of NNs vs traditional methods.

Historically – tabular data predates image, text + audio data – focus of ‘early ML research’ – NNs developed mainly on homogeneous data.

Reasons for GBDTs outperforming NNs ‘unclear’.

‘Low-Quality Training Data’ – tabular data – ‘often include missing values, extreme data (outliers), and erroneous or inconsistent data + small overall size relative to the high-dimensional feature vectors generated form the data.’

Also issues with class imbalance – *Class imbalance between goal and not goal in shots*.

These issues face GBDTs however – able to adjust w ‘appropriate approximations and split values.’

‘Missing or Complex Irregular Spatial Dependencies’ – ‘often no spatial correlation between the variables in tabular datasets or the dependencies between features are rather complex and irregular.’ – ‘Relationships between features have to be learned from scratch’ – ‘Thus, the inductive biases used in popular models for homogenous data, such as CNNs, are unsuitable for modelling this data type.’

‘Dependency on Preprocessing’ – in homogenous data preprocessing not that important or necessary – however – for tabular data – preprocessing very important – esp dealing with categorical features – one hot encoding leading to sparse feature matrix – these steps have potential to cause ‘information loss’ – lower performance.

‘Importance of Single Features’ – GBDTs much more adept at understanding varying feature importance – NNs less so – data like image, audio etc much more even feature weight.

Possibly using different methods for cat variables – leave one out encoding – replace value with mean of target variable.

Possibly using transformers.

More continuous features better for NN – very large datasets (in the millions of rows) also better for NN.

Gradient boosting ‘over 20 years’ old – XGBoost first published in 2017 – still remain highest performing.

Kadra

Tabular data – ‘understudied’ with reference to NNs.

Regularisation key to improving NNs.

Methods for regularisation: ‘Weight decay’, ‘Data Augmentation’, ‘Model Averaging’, ‘Structural and Linearisation’, ‘Implicit’.

Fayaz (2022)

Ensemble method best.

NNs highly successful in ‘audio, images and text data.’

NN issues: ‘missing data, mixed data (nominal, numerical and categorical), data imbalance, data overfitting, and a lack of specific knowledge about the dataset’s structure.’

Used Hyperopt – 1000 steps.

Flepp (2020)

‘Match outcomes in football are disproportionately influenced by randomness because football is a low-scoring game in which winning and losing is often determined by a single goal.’ – ‘Thus, match results occasionally fail to reflect the true level of play of the two teams on the pitch’ – ‘it is questionable whether match outcomes truly represent a reliable performance indicator, particularly when considering a limited match window within the scope of a single season’.

‘Outcome-based performance evaluation’ – leads ‘systematic misjudgement’.

Hard for people to accept randomness – ‘people mistakenly perceive patterns in random sequences.’

Potential for ‘outcome bias’ – ‘role of randomness in match outcomes’ underestimated – ‘assign too much weight to the observed outcomes in their performance evaluation’.

‘Decision makers fail to make needed adjustments after fortuitous wins and act excessively after unlucky losses’.

Expected goals – quantifying the quality of chances created (by probability of a chance leading to a goal) – better reflection of the teams performance.

‘Scoring chances are intrinsically tied to scoring goals, it can be assumed that they carry informative signals’.

Benefits:

More ‘scoring chances’ than goals – less randomness involved – *weak law of large numbers.*

Accounts for quality of chance – acknowledges not all shots are equal value.

‘Intuitive’ for football context – minimising the opponents ‘scoring chances’ and maximising your 'scoring chance’ is an idea consistent with football thinking.

Can be applied to team performance with metrics such as expected goal difference.

Expected goal difference a better predictor of future results than previous match outcomes (given by higher R2 value in a univariate linear regression comparing xGD and previous points on future points) – particularly effective in small sample sizes where effects of randomness are greater.

*This research did not have access to opposition and player teammate data which I do.*

In most xG models – the finishing and goalkeeper skills are assumed to be average.

General idea to focus on things which are in the control of the players.

R2 for xGD of prev 10 matches on points for next 10 0.32 compared to 0.25 for points won.

Location of shot key variable.

Rathke (2017)

Precursor analysis found a relationship between successful teams and teams which took a lot of shots.

Location of shot key variable.