Project Plan

Literature Review

Deep learning and neural networks have been highly successful in many applications, particularly those focusing on homogenous data such as image, text and audio (Armon). However, in the case of heterogenous data such as tabular data, more traditional methods particularly gradient boosted decision trees (GBDTs) still outperform neural networks in most cases (Armon). This underperformance in tabular data is especially important due to the widespread use of tabular data making it the most common dataset type, with applications in a variety of fields such as medicine, climate research and finance (Armon). Tabular data is also older than the newer audio, text and image data which neural networks perform best on, with traditional methods being focused on tabular data compared to more recent methods such as neural networks. The persistence of superiority for GBDTs is especially remarkable as gradient boosting was first developed more than 20 years ago and one of the most successful GBDT implementations XGBoost was first published in 2017. Thus, it is clear that the performance of neural networks on tabular data is a phenomenon that is unique and warrants investigation.

Common issues which face neural networks when working with tabular data are the mixed data types for features, potential for missing values and the inability for a model to be pretrained with prior knowledge like in text and image applications (Armon). Another problem for neural networks is the presence of uninformative features in many tabular datasets and an irregular distribution of feature importance (Grinsztajn). This is in contrast to text and image analysis which have more equally balanced feature importance (Grinsztajn). Neural networks are also less able to handle class imbalances compared to GBDTs which is a common feature amongst tabular datasets. As well as this the correlation and relationships between features is frequently complex or there is no correlation at all in tabular data, this is another means of difficulty which neural networks face in tabular data compared to other data types. Furthermore, in comparison to audio, text and image data, tabular data requires more preprocessing, especially for categorical features where one hot encoding can lead to a sparse feature matrix with the preprocessing steps potentially hindering performance due to information loss.

Investigations into the characteristics of tabular datasets associated with superior performance for neural networks have yielded mixed results. Most clearly in dataset size where the study of Borisov et al suggests that very large datasets including millions of observations are best for neural networks, however, McElfresh et al have suggested that neural networks perform better in smaller datasets. Other characteristics of tabular datasets where neural networks are successful include datasets where features are distributed normally with regular tails, as well as datasets with predominantly continuous variables.

Potential improvements to neural networks in relation to tabular data have been suggested in recent years with architectures specifically designed for neural networks such as TabNet. Other suggestions include regularisation methods, improving handling of categorical variables and transformer approaches. As well as this to reduce the impact of uninformative features, removing the least valuable features is a potential step to improving neural networks.

I aim to investigate neural network performance on tabular data relative to GBDTs specifically focusing on the prediction of expected goals in football analysis. Football is a low-scoring sport and due to this matches are often only decided by one goal resulting in randomness have a large impact on match outcomes. Due to this match outcomes often not necessarily reflecting the actual performance levels of the teams involved, meaning that match results are not an ideal predictor of performance especially in small sample sizes. An alternative is expected goals, which quantifies the quality of chances created, given by the sum of the probability of each shot resulting in a goal. This represents an improvement as there are more shots in a game than goals, increasing sample size and reducing randomness. Moreover, by predicting the probability of each shot resulting in goal the quality of each shot is accounted for recognising that not all shots have equal value. Investigations into the predictive value of expected goals compared to match results (points), have shown a higher R-squared value for using expected goals to predict future match results (points) in a univariate linear regression compared to using previous match results to predict future match results.

Based on the literature I would expect that this context may be more suited to GBDTs compared to neural networks. This is because there is a class imbalance between goals and not goals, in the dataset. Moreover, there are mixed data types, between categorical, coordinate and numeric, which also may prove challenging for neural networks. It is also likely that the feature importance for the variables in the dataset will vary, with it being suggested that the distance from the goal and the angle to the goal being the most important features.