Mid-Term Report: NFL Big Data Bowl 2024

Han Zhang, Kefan Jiang, Leonardo Zhu, Tinhang Hong

April 2024

- **Which algorithms and baselines have you implemented?**
  We implemented **lightGBM** and **neural networks** models to predict the play results of football games and **mean absolute error** to measure the accuracy of models.

- **What challenges did you encounter? How did you solve it or how are you planning on solving it?**
  Feature Engineering:The dataset includes many features that it takes time to think about how to select those useful features. It requires domain knowledge. To solve the challenge, we need to have a deeper understanding of football rules and watch many football games. We also did some common feature engineering tools like standardization.
  Merging dataset: Our dataset incorporates many different sub datasets and lots of the data are "dirty". We manage to merge all of the datasets into one datasets through pandas. We also extracted and created some important features through raw datasets. Interpretability: Tree-based models like LightGBM can be challenging to interpret compared to simpler models, especially when they involve a large number of trees or complex tree structures. Neural networks are often considered "black box" models, meaning it can be difficult to understand how decisions are being made, which is a significant challenge in fields requiring explainability.

- **What would you like to complete by final report? What are additional experiments you would like to include if there is time?**
  The only step left for us to do is hyper parameter tuning through k fold cross validation on neural networks. We are considering to tune the best hyper parameters for number of layers, regularizations and activation function, etc. We want to look at the validation loss at every models' with different hyper parameters and choose the one that has the best performance. We are also considering to use categorical embedder to embed the categorical features.

- **Discuss preliminary results and analysis if available or analysis proposed.**
  We fitted the distribution of our label and found out it is normally distributed. If we directly fit a normal distribution based on the sample

mean and sample variance, the MAE we get is around 10. While our lightgbm has an error of 5.1 and neural networks has an error of 4.4, our models and experiments turn out to be pretty successful. We also conducted correlation analysis to detect correlated features and performed feature engineering. We also used label encoder to encode the categorical features and it turned out that label encoder was a better tool then one hot encoder. Perhaps our features have some ordinal information among them.

- **Any changes in scope or project direction?**
  We changed some analysis tools. The final goal changed from predicting **"HomeFinalScore"** to predicting **"PlayResult"**(Net yards gained by the offense). We also implemented auxillary loss in neural networks by adding passLength as an auxillary label to help us predict the results.