Project Assignment 1 (20 pts)

10-718: Machine Learning in Practice

Due Thursday, February 8 at 9:30am ET

Instructions: Provide answers to the questions below and submit your writeup via Gradescope. When submitting your writeup on Gradescope, please assign pages to each question correctly (it prompts you to do this after submitting your work), which will help to streamline the grading process.

1 Problem Formulation [6 pts]

Provide a detailed problem formulation for the dataset you've selected, including the following items:

a) Model Goal (2pts): What problem do you aim to solve with your selected dataset? Are you solving a prediction (e.g., classification, regression) or generation task? What is the intended output of the model you will develop (e.g., 0/1 labels, probabilities, numerical values, images, text)?

I aim to solve the problem in football of identifying possible blitzing defenders before the play happens based on defensive presnap position. By predicting which players on the defense will run at the quarterback, the offense can effectively counteract and protect the safety of their quarterback. This is a prediction task; given the tracking data of each player on the field before a play happens, I want to output whether any given player on defense will blitz the quarterback. The intended output of the model I will develop will be 0/1 labels for each defensive, indicating a blitz, and corresponding probabilities of them blitzing.

b) System Goal (2pts): Imagine that you are using the model you develop as part of a larger system/application. Give an example of one system/application that could benefit from your model.

There are a multitude of applications that could benefit from this accurate blitz prediction model. Amazon currently uses a blitz prediction model in their national broadcast of NFL games to highlight the potential blitzers on the screen: Athletic Article. One application that could benefit from an open source blitz prediction presnap model would be the widespread usage of the blitz identification on all NFL broadcasts, outside of the ones done by Amazon. This could help broadcasters and announcers provide live-time insights which can improve viewership and interest in the sport.

c) Measuring Success (2pts): How will you measure the effectiveness of your model? Provide at least two measurements that could be used to evaluate the efficacy of the model. Additionally, suppose you use the model in the application you described in part (b). How might you measure success of the overall system that uses the model?

A basic evaluation of this binary classification problem would include accuracy, precision, recall, and F1 score. On the probability scale, I will also use cross entropy loss to asses the model's outputs. If I was to measure the success of adding the live blitz prediction model to NFL broadcasts, I could test for its success by analyzing viewer retention and conducting viewer surveys on the addition of the blitz model. I could also evaluate sentiment on social media of the new model - the success rate would be based upon human feedback.

2 Dataset Exploration [8 pts]

a) Visualization (3pts): Provide at least one visualization/summary of your dataset (e.g., summary statistics, box plots, histograms, correlation matrix).

The Big Data Bowl 23 dataset comes with the following data.

- **games.csv**: Game level information about the date and teams playing. 122 games total from the 2021-2022 NFL season.
- **plays.csv**: Play level information for each play in every game of the dataset. Consists of the result, formations, and variables describing what happened on the play. 8557 plays from the 2021-2022 NFL season.
- **players.csv**: Player level information describing the background and biometrics of a given player. 1679 players frrom the 2021-2022 NFL season.
- **pffScoutingData.csv**: Scouting data done by the company PFF analyzing every player on the field for a given play. Provides the player's role (including whether they blitzed or not) along with a grade of the player's actions. 8557 plays * 22 players per play = 188254 data points.
- **week[week].csv**: Tracking level data for every play from week [week]. Each play has multiple frames captured, and 5 captured before the ball is snapped. We are interested in the player's position before the ball is snapped.

The tracking level data denotes x,y for each plater on the field. See Figure 1 for x,y positioning on the field.

- x is the player position along the long axis of the field, 0 - 120 yards.
- y is the player position along the short axis of the field, 0 - 53.3 yards.

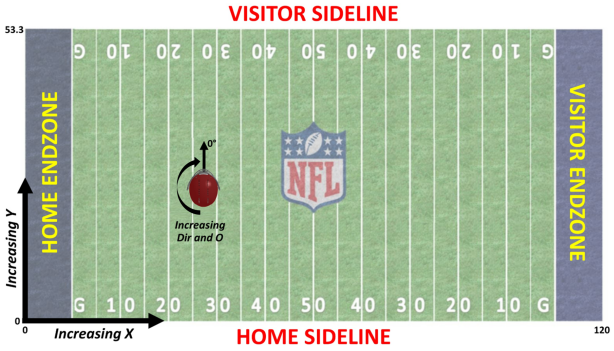

Figure 1: Tracking Player x,y coordinates

For each play, both offense and defense line up in a set personnel. The personnel of each can predicate which players on defense will blitz. In Figures 2 and 3, I display the breakdown of the top 5 most common personnel across all the plays in the 2021-2022 NFL season.

Our outcome for each player we are interested in is provided in the PFF dataset, where they mark each player from every play with a label of whether they blitzed or not. The **mean** number of blitzers across all the plays in the dataset was **4.25**. The **median** number of blizters across all plays was **4**. The **minimum** blitzers was **1**, and the **maximimum** blitzers was **8**.

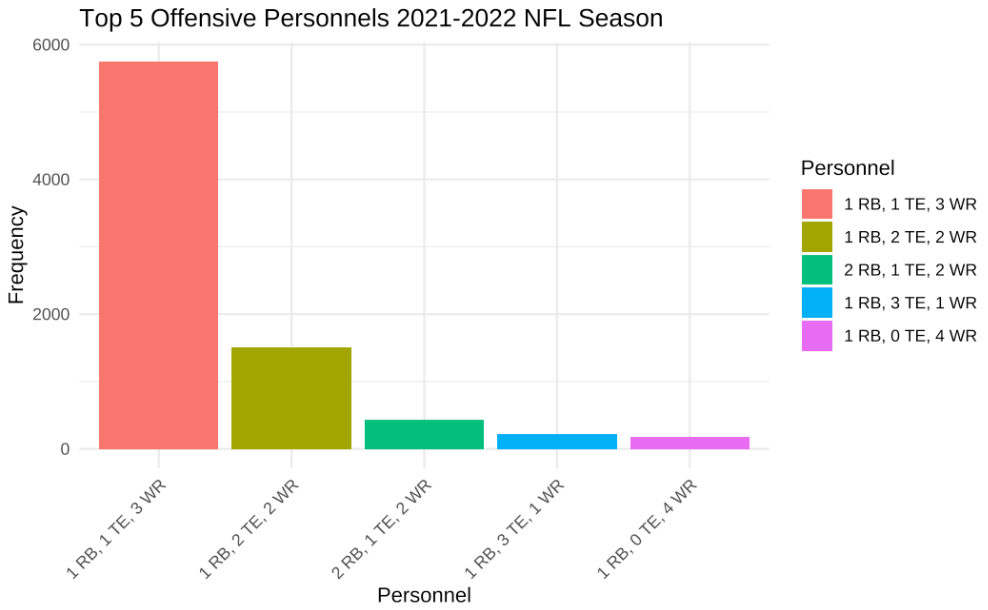

Figure 2: Top Offensive Personnels Used

b) Cleaning (2pts): What data cleaning or preparation is necessary for your data? Provide at least one example of cleaning/preprocessing you could perform on the data and explain why it might be helpful.

One large data cleaning task I have is to join the datasets. Since the plays, players, and scouting data are all in separate data tables, I have to join these tables by unique keys to get all the information in one table. Outside of joining tables, I anticipate having to encode a lot of string variables as factor variables - this will be necessary for modeling outcomes and using string variables as predictors. This will allow for more efficient memory and manually ordering of factoring models.

c) Feature Analysis (3pts): Consider what features might be useful for your selected dataset and problem. Describe the features and perform at least one analysis to explore the utility of these features.

I anticipate the grouping of players on the field will be inportant when predicting if they are blitzing. For example, outside of the coordinates, a player's assigned position will affect their probability to blitz. Additionally, game conditions can also affect the probability of a player blitzing. If a certain state of the game favors an aggressive approach on the defensive side, this can increase the chance of players blitzing. Let's look at the average number of blitzers per play, grouped by player position and game condition. When looking at game condition, let's start by simply considering the down of the play. A down signifies a period a play occurs, every team has 4 downs to try and achieve a first down, which is attained by advancing ten yards. We see in Figure 4 and 5 the average number of rushers in a play across downs and defensive positions. There seems to be little effect captured by the down, and some correlation with defensive position and higher rates of rushing. This data allows for a lot of grouping, so I anticipate using multilevel modeling as a baseline.

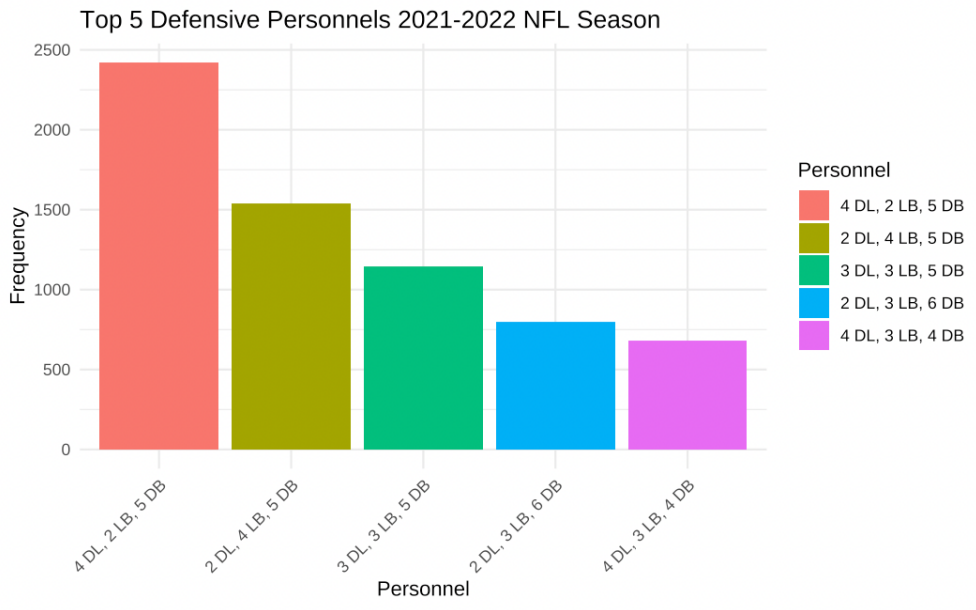

Figure 3: Top Defensive Personnels Used

3 Pipeline Sketching [6 pts]

Propose a **detailed pipeline** to solve the machine learning problem you identified in Section 1 (e.g., including data cleaning/preprocessing steps, data/feature engineering strategies, models you'd like to test, evaluation techniques, deployment considerations). **What is a simple baseline you will test?**

Note: You do not have to train any models or perform any steps of the pipeline for this question, but preliminary analysis is encouraged. The main purpose is for you to understand the problem and formulate a roadmap for the analysis you want to conduct in the subsequent assignments.

3.1 Data Cleaning and Preprocessing Steps

There are numerous data cleaning steps needed to prepare the data for analysis. First, I plan on merging the data sheets given by the Big Data Bowl. This will allow for a unified dataset, with every player for every play on a given play's tracking data with player, team, and scouting insights. Additionally, I will find NAs and outliers by inspecting the distributions of interested variables, removing these observations or using data imputation if more appropriate. I plan on also creating a train/test set.

3.2 Feature Engineering Steps

Once I have cleaned my data, I will move to manipulating the variables in the dataset. Encoding string variables as ordered factor variables will be one of the first tasks to do. I will also look into normalizing variables if necessary. Additionally, I plan on computing intermediate statistics on the player and team level, such as average positioning, blitz rate, and relevant metrics that can be drawn from the data. Since the tracking data is captured over many frames, I also plan on creating variables that consolidate this information for non-sequential models. For example, if I want to capture how fast or far a player has been moving across 5 frames, I can utilize the speed and coordinated captured across all 5 frames to produce a summarized

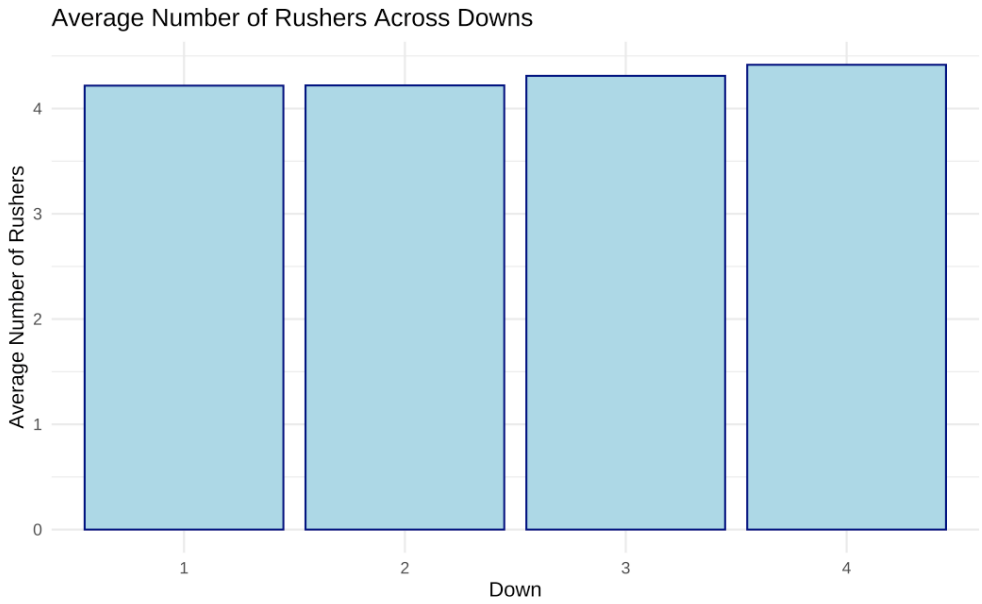

Figure 4: Average Number of Blitzers per Down

metric. This should help capture contextual metrics that would not be captured without incorporating sequential calculations. I plan on also creating tabular representations of each player on the field to use for deep learning models.

3.3 Model Baselines

I plan on implementing a couple simple classification baselines. These include a Naive Bayes Classifier, Logistic Regression, and Random Forests. Simply given presnap information at one time point (non-sequential modeling), I will model each player's probability/label of blitzing. This will include mostly variable selection and using these baseline models to investigate which variables are most important towards predicting blitzing. These models will be evaluated using the metrics mentioned in the earlier section, such as accuracy, precision, recall, and F1 score.

3.4 Multilevel Models and Deep Learning Techniques

After implementing our baselines and creating a standard for prediction accuracy, we will move to more complex models. Using the takeaways of variable importance from our baselines, I will try different random and fixed effects in multilevel models in this classification problem to utilize the aformentioned grouping stucture of our data. Additionally, I also plan on using sequential models on the tracking data to incorporate the frames given. Using RNNs or CNNs, I plan on training smaller deep learning models to predict the blitzers given the tabular position data pproduced in the feature engineering task.

3.5 Evaluation and Model Selection

Given all the models are classification models, we can you the evaluation metrics mentioned before. Comparing across models, we will find the best model and analyze the differences between all of our complex methods and baselines. One important analysis will be failure modes - what does multilevel modeling capture that

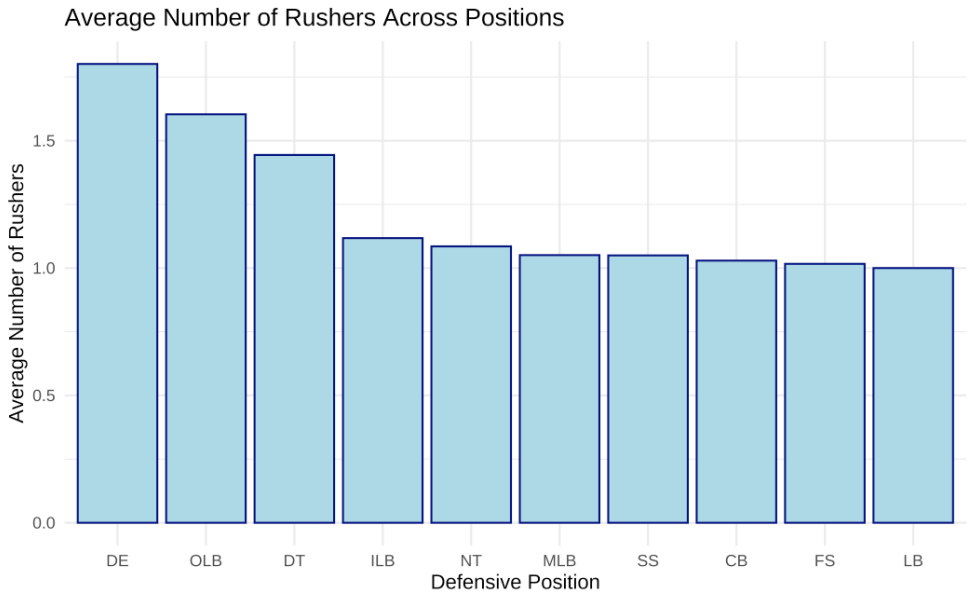

Figure 5: Average Blitzers on Play per Position

sequential modeling struggles with? Do the baselines perform better than the complex methods at certain situations? Do they perform better in general.

3.6 Releasing Analysis and Code

After writing up my analysis and storing the code for all of my models and data preparation, I plan on making my work public to encourage comparison with the Amazon model and adaptation to other broadcasts.