SRS_CIS-519-401 2021A 2-Page Project Proposal

Jonathan Yushuvayev, Hannah Kronenberg

TOTAL POINTS

12 / 15

QUESTION 1

1 Completed Template 12 / 15

- **0 pts** All components
- **1 pts** No Team List
- **3 pts** No clear formulation
- **2 pts** Work plan
- **3 pts** Motivation, methods, challenges
- **1 pts** Prior relevant work
- **1 pts** Incorrect/inconsistent project title.
- **0.5 pts** Partial credit for prior relevant work.

✓ **- 1 pts Partial credit for prior relevant work.**

- **0.5 pts** Partial credit for Task, Experience or Performance.

✓ **- 1 pts Partial credit for Task, Experience or Performance.**

- **2 pts** Partial credit for Task, Experience or Performance.

✓ **- 0.5 pts Partial credit for Methods and Results.**

- **1 pts** Partial credit for Methods and Results.

✓ **- 0.5 pts Partial credit for Remaining Challenges and Work plan.**

- **1 pts** Partial credit for Remaining Challenges and Work plan.

Some remarks are listed below. Please read in detail and feel free to discuss with me during cohort meetings or OHs.

On Related Prior Work, since it is a Kaggle competition, you should add more details on the existing literature and how the problem is approached. You should also describe some of the more successful methods within the competition.

On Formal Problem Setup, under Task, you should describe the machine learning task, rather than describing concepts related to the dataset only. In Experience (not Experiments), you should also briefly describe the planned methodology (how you are doing preprocessing, what models you plan to use, how you plan to train the models), rather than describing the dataset only.

On Cleaning: In general, try to give more context and details behind the data preprocessing steps, as they are highly dependent on the interpretation of the algorithm designers. For instance, you could give some intuition or explain why it is necessary to “reshape the dataset to one row per play”. Also, “dropping out 10 plays which had missing velocities” is fine, but you should justify the “small number of plays” by providing a ratio of those plays against the total number of plays. In general, write the report in a way such that it provides sufficient background to the reader and right now, it is a little lacking in this aspect.

On Modeling and Results: In subsequent reports, try to give more details on the machine learning tools you are using. As a general guideline, you should write it in such a way that if anyone reads your report, he/she will be able to reproduce all, if not most of the results. For instance, it appears that you are using a random forest regressor and an adaboost regressor with a decision tree regressor. You should mention these in more detail, including all the parameters you used. Also, there is something

fundamentally unsound about the way you describe Adaboost. Adaboost is an algorithm that can enhance an existing regressor and hence, it has to be used together with a regressor. In your description, there is no mention of any regressor that is used in conjunction with Adaboost, which makes it almost impossible for anyone to figure out how modeling and training is done.

You should also expand on your cross-validation process (how many folds, size of data sets etc.) and how your grid search of hyperparameters is carried out. I understand that there is a 2-page limit for this progress report, but you should include these details in subsequent reports. Also, apart from your cross-validation and test errors, you should also report your training errors. Your training errors are the first indication of whether your model works well or not. When training with neural networks, you should also report and analyze the training, validation/test losses, together with those accuracies mentioned above. Since this is a Kaggle competition, you should benchmark your results against some of the results on the leaderboard to see how well you perform.

On Remaining Challenges and Plan: For your plan on feature engineering, it's important that you cite any of the ideas you have gotten elsewhere (see note on plagiarism below). It is mentioned in your Week 2 plan that you will "try a more basic statistical method, like regression". From the section on Modeling and Results, it appears that you are already using regression for your random forest and adaboost+decision tree approaches, so what is the difference between what you have done versus what you plan to do in Week 2? For Week 3, before proceeding to use a neural network, try to understand which are the more important features and that should help you in improving performance, for both existing and new approaches.

General comment: In general, my grading criteria will not be solely based on the performance of your model or the results alone. If you have performed a thorough analysis and pre-processing of the data based on your own understanding and ideas, implemented correct ML models and analyze the strengths and weaknesses of these models with regards to your problem setup, described your modeling, training and validation methodology in an accurate and detailed manner, then you are in good shape.

Important note on plagiarism: Since this is a Kaggle competition with extensive documentation and open-source code, you will need to pay special attention to this and make sure that you perform your own code implementation, analysis and data preprocessing such that it does not plagiarise existing materials. When in doubt, you should cite any references, including suggestions on data processing or analysis.

Predicting Yards Ran Based on Detailed Game Data

Team: Hannah Kronenberg (519, KongYao's Cohort), Jonathan Yushuvayev (519, KongYao's Cohort)
Project Mentor TA: Kong Yao Chee

Introduction

For this project we are attempting to predict the number of yards gained on a running play of an american football game based on the state of the game at the time the ball is handed off. This problem is interesting for NFL teams because it could help them make better decisions around run plays. Furthermore, it is a problem that is of interest to our team despite not being a particularly "important" problem. The data and problem statement was taken from the Kaggle competition "NFL Big Data Bowl."

Related Prior Work

Winner of the NFL Big Data Bowl Competition:
Because this is a completed Kaggle Competition we know that the winner used a Convolutional Neural Network Approach (CNN) approach. Comments from the competition were also helpful in the data cleaning process. In particular, there were notes about variables that were not reliable.

Formal Problem Setup (T, E, P)

Task T:
For play p, the state of the game at handoff is given by twenty-two 4-dimensional vectors corresponding to the position and velocity in x,y coordinates of each player on the field. The players in the defending team and the teammates of the ball carrier are interchangeable so we could in principle use data augmentation.
Experiment E:
We are using a dataset of running plays during the 2017-2019 NFL seasons that specifies the players involved in each play with their position at the time of the handoff as well as the teams, scores, time during the game that the play happened in, as well as the number of yards gained.
Performance P:
We use Mean Squared Error (MSE) to measure the accuracy of our predictions over the held out dataset.

Progress Report (Methods, Results):

Cleaning
We start by bringing in the data and performing some basic cleaning. This includes:

- Standardizing team names across columns
- Standardizing the direction of all of the plays in the dataset
- Calculating velocities of each player in the x and y directions

- Reshaping the dataset so that we have one row per play as opposed to one row per play per player (which is how the data comes in).
- Dropping around 10 plays which had missing velocities. We could potentially replace those velocities with 0 but since it is a small number of plays we thought it would be fine to drop for now.

Modeling

- Our first attempts at modeling were to use a random forest and adaboost with a decision tree.
- We performed grid search cross validation to tune the number of trees in both methods.
    - In the random forest we did the number of features at each split.
    - In adaboost we did the depth of the tree.

Results

Random Forest

n_estimators	max_features	avg_score	score_std_dev
30	0.3	42.163014018	4.341433988
30	0.6	42.672040060	3.890607655
30	1.0	42.778728508	4.074871041
60	0.3	41.574686657	4.087079751
60	0.6	41.948111956	4.156407229
60	1.0	41.768485829	4.026741706
100	0.3	41.145994479	4.101245519
100	0.6	41.409153686	4.127515936
100	1.0	41.703929722	4.182286437

Adaboost

n_estimators	max_depth	avg_score	score_stddev
20	2	87.625044353	28.199625309
20	3	109.63393861	9.2758083027
20	4	93.552030772	4.0068172836
40	2	88.283386986	22.487162921
40	3	126.17472126	27.767063045
40	4	127.53979483	6.7285422325
60	2	80.561290975	36.841757167
60	3	148.10050460	19.401574476
60	4	161.47941731	5.6419317119

Final

Method Name	CV MSE	Test MSE
Random Forest	**41.15**	42.03
Adaboost	**80.56**	59.84

Remaining Challenges, and work plan over the next ~4 weeks:

- Remaining Challenges: build a neural network, investigate high MSEs, feature engineering (to be done over the next four weeks)

1. **Week #1 (3/28-4/3):** Our mean squared errors are really high - this week we will investigate if there is a problem with the way we are fitting our models
2. **Week #2 (4/4-4/10):**
    a. Feature engineering to improve performance
        i. sort the defenders and attackers by distance to the football
        ii. Try adding distance to the closest defender as a feature
    b. Try a more basic statistical method, like regression
3. **Week #3 (4/11-4/17):** try to build a neural network (leaning to convolutional)
4. **Week #4 (4/18-4/24):** fix any bugs, compare the models, write up the results

Prior Work / References (this part can be over the 1-pg limit):

1. NFL Big data Bowl (https://www.kaggle.com/c/nfl-big-data-bowl-2020)
2. Winning Submission Post (https://www.kaggle.com/c/nfl-big-data-bowl-2020/discussion/119400)
3. Fernandez Javier, Bornn Luke, “Wide Open Spaces: A statistical technique for measuring space creation in professional soccer” http://www.lukebornn.com/papers/fernandez_ssac_2018.pdf

1 Completed Template 12 / 15

- **0 pts** All components
- **1 pts** No Team List
- **3 pts** No clear formulation
- **2 pts** Work plan
- **3 pts** Motivation, methods, challenges
- **1 pts** Prior relevant work
- **1 pts** Incorrect/inconsistent project title.
- **0.5 pts** Partial credit for prior relevant work.

✓ **- 1 pts Partial credit for prior relevant work.**

- **0.5 pts** Partial credit for Task, Experience or Performance.

✓ **- 1 pts Partial credit for Task, Experience or Performance.**

- **2 pts** Partial credit for Task, Experience or Performance.

✓ **- 0.5 pts Partial credit for Methods and Results.**

- **1 pts** Partial credit for Methods and Results.

✓ **- 0.5 pts Partial credit for Remaining Challenges and Work plan.**

- **1 pts** Partial credit for Remaining Challenges and Work plan.

Some remarks are listed below. Please read in detail and feel free to discuss with me during cohort meetings or OHs.

On Related Prior Work, since it is a Kaggle competition, you should add more details on the existing literature and how the problem is approached. You should also describe some of the more successful methods within the competition.

On Formal Problem Setup, under Task, you should describe the machine learning task, rather than describing concepts related to the dataset only. In Experience (not Experiments), you should also briefly describe the planned methodology (how you are doing preprocessing, what models you plan to use, how you plan to train the models), rather than describing the dataset only.

On Cleaning: In general, try to give more context and details behind the data preprocessing steps, as they are highly dependent on the interpretation of the algorithm designers. For instance, you could give some intuition or explain why it is necessary to “reshape the dataset to one row per play”. Also, “dropping out 10 plays which had missing velocities” is fine, but you should justify the “small number of plays” by providing a ratio of those plays against the total number of plays. In general, write the report in a way such that it provides sufficient background to the reader and right now, it is a little lacking in this aspect.

On Modeling and Results: In subsequent reports, try to give more details on the machine learning tools you are using. As a general guideline, you should write it in such a way that if anyone reads your report, he/she will be able to reproduce all, if not most of the results. For instance, it appears that you are using a random forest regressor and an adaboost regressor with a decision tree regressor. You should mention these in more detail, including all the parameters you used. Also, there is something fundamentally unsound about

the way you describe Adaboost. Adaboost is an algorithm that can enhance an existing regressor and hence, it has to be used together with a regressor. In your description, there is no mention of any regressor that is used in conjunction with Adaboost, which makes it almost impossible for anyone to figure out how modeling and training is done.

You should also expand on your cross-validation process (how many folds, size of data sets etc.) and how your grid search of hyperparameters is carried out. I understand that there is a 2-page limit for this progress report, but you should include these details in subsequent reports. Also, apart from your cross-validation and test errors, you should also report your training errors. Your training errors are the first indication of whether your model works well or not. When training with neural networks, you should also report and analyze the training, validation/test losses, together with those accuracies mentioned above. Since this is a Kaggle competition, you should benchmark your results against some of the results on the leaderboard to see how well you perform.

On Remaining Challenges and Plan: For your plan on feature engineering, it's important that you cite any of the ideas you have gotten elsewhere (see note on plagiarism below). It is mentioned in your Week 2 plan that you will "try a more basic statistical method, like regression". From the section on Modeling and Results, it appears that you are already using regression for your random forest and adaboost+decision tree approaches, so what is the difference between what you have done versus what you plan to do in Week 2? For Week 3, before proceeding to use a neural network, try to understand which are the more important features and that should help you in improving performance, for both existing and new approaches.

General comment: In general, my grading criteria will not be solely based on the performance of your model or the results alone. If you have performed a thorough analysis and pre-processing of the data based on your own understanding and ideas, implemented correct ML models and analyze the strengths and weaknesses of these models with regards to your problem setup, described your modeling, training and validation methodology in an accurate and detailed manner, then you are in good shape.

Important note on plagiarism: Since this is a Kaggle competition with extensive documentation and open-source code, you will need to pay special attention to this and make sure that you perform your own code implementation, analysis and data preprocessing such that it does not plagiarise existing materials. When in doubt, you should cite any references, including suggestions on data processing or analysis.