Capstone: NFL Draft Machine Learning Model

Western Governors University

001501081

Jeremiah Hennessy

12/13/2021

Table of Contents

[**Prompt A** 4](#_Toc90474534)

[**Letter of Transmittal** 4](#_Toc90474535)

[**Project Recommendation** 7](#_Toc90474536)

[**Application Benefits** 8](#_Toc90474537)

[**Objective and Hypothesis** 9](#_Toc90474538)

[**Methodology** 10](#_Toc90474539)

[**Funding Requirements** 11](#_Toc90474540)

[**Stakeholders Impact** 11](#_Toc90474541)

[**Data Precautions** 11](#_Toc90474542)

[**Developer Expertise** 12](#_Toc90474543)

[**Prompt B** 13](#_Toc90474544)

[**Project Proposal** 13](#_Toc90474545)

[**Data** 14](#_Toc90474546)

[**Project Methodology** 16](#_Toc90474547)

[**Project Outcomes** 17](#_Toc90474548)

[**Implementation Plan** 17](#_Toc90474549)

[**Evaluation Plan** 18](#_Toc90474550)

[**Resources And Costs** 19](#_Toc90474551)

[**Timeline And Milestones** 19](#_Toc90474552)

[**Prompt C** 21](#_Toc90474553)

[**Application Files** 21](#_Toc90474554)

[**Prompt D** 22](#_Toc90474555)

[**Post-implementation Report** 22](#_Toc90474556)

[**Project purpose** 22](#_Toc90474557)

[**Datasets** 26](#_Toc90474558)

[**Data product code** 30](#_Toc90474559)

[**Hypothesis verification** 31](#_Toc90474560)

[**Effective visualizations and reporting** 32](#_Toc90474561)

[**Accuracy analysis** 33](#_Toc90474562)

[**Application testing** 34](#_Toc90474563)

[**Appendices** 35](#_Toc90474564)

[**Installation Guide** 35](#_Toc90474565)

[**User Guide** 36](#_Toc90474566)

[**Summation of Learning Experience** 37](#_Toc90474567)

# **Prompt A**

## **Letter of Transmittal**

December 3rd, 2021

Mr. Jimmy Haslem

Cleveland Browns

100 Alfred Lerner Way

Cleveland, OH 44114

Mr. Haslem,

The NFL draft is the most crucial way that NFL teams can find cheap talent that will lay a foundation for that organization in years to come. Organizations spend thousands of hours studying collegiate talent to find the next generation of NFL stars. Traditional methods of evaluating these prospects include in person scouting, interviews, combine reviews, and actual college football tape review. When draft day comes around, the grade your organization has on players is just as important as the grade that other teams have on players. Understanding the value that the league has on a player helps organizations understand appropriate moments to trade up or down in the draft, and appropriate moments to reach for a player that so called “experts” have associated with a low grade. League gossip and media give a good understanding of how teams value players, but it is possible to gain more insight by using past draft data.

Every year NFL prospects are tested rigorously in different areas at the NFL combine. The test results of these prospects have an immense impact on which round these players will be selected in come draft day. By collecting all the data for past draft prospects and analyzing this data, it is possible to build a machine learning model that will be capable of predicting the round that a player will be drafted in based on previous combine performances. The model will correlate which combinations of test results lead to which round (if any) the players will be drafted in. Experts give a grade for each player’s combine performance each year. Using this tool to evaluate the combine performance of players will give historical context to the combine grade that each player deserves, instead of basing it solely off opinion.

The cost to develop this model will be an upfront cost of $50,000, and a yearly fee of $100,000 for continual use. Maintenance and upgrades will be included throughout the duration of the contract between the Cleveland Browns and my company.

The implementation of this project will rely upon knowledge obtained from my Bachelor of Computer Science along with deep self-study of machine learning and data science. I have also been an avid NFL fan and have spent around 10 years of my life studying the process of prospects being drafted into the NFL.

Sincerely,

Jeremiah Hennessy

## **Project Recommendation**

**Problem Summary**

Current methods of evaluating combine performances of NFL draft prospects rely on the opinions of expert scouts. While these scouts have a deep knowledge of football and the measurables that will lead to NFL draft value, a greater understanding can be made using a deep analysis of past combine performances. It is possible to use machine learning to help make a concrete and factual analysis of where NFL prospects will be drafted. Through combining expert opinions and a historically based statistical analysis, the Cleveland Browns will have the greatest advantage in evaluating the value of NFL draft prospects.

The solution that will be created will use machine learning to create models based on past draft data and combine data. The models will be trained to accept the combine statistics from a draft prospect and make a prediction on which round that player will be drafted based on their combine performance.

### **Application Benefits**

This product will give the Cleveland Browns a greater understanding of where players will be drafted on draft day. Understanding where a player will be drafted is crucial in building an NFL draft strategy. It is important to know when it is appropriate to wait on selecting a player and to know when it is appropriate to select a player early. This also plays into deciding when it is appropriate to trade draft picks to move up or down in the draft.

**Application Description**

The application will be a dashboard that walks through the process of analyzing past combine performance data. The dashboard will manipulate and dissect the data and then use the data to create machine learning models that will be capable of guessing where players will be drafted.

**Data Description**

The data used for this application will be collected from the NFL Play Statistics Dataset on Kaggle.com. The application will use two csv files. One file will be a list of NFL draft data from 1987 to 2019, and the other file will be a list of combine performance of every prospect from 1987 to 2019. These files will be combined in the application using the combine ID given to each prospect. The data will be mostly numerical, but college information data and position data will be categorical. Certain positions will be excluded from the dataset because of the lack of combine performance correlation to draft value. Prospects that attended the combine and were not drafted will still be included in the data to help the models understand combine performances that lead to not being drafted. Each data point collected at the combine will be treated as an independent variable and the round prospects were drafted in will be treated as the dependent variable. Today the NFL draft contains 7 rounds. The data used for this application will includes years in which the NFL draft was 11 rounds. Any prospects drafted after round 7 will be treated as an undrafted player to comply with current NFL draft format.

### **Objective and Hypothesis**

The objective of this project is to create a dashboard that successfully analyzes past combine performances in relation to draft value, and that successfully uses machine learning to create models capable of predicting where prospects will be drafted based on their combine performance. The final product is expected to be built under the premise that combine performance will influence which round prospects will be drafted. If enough combine data and NFL draft data are correctly formatted and manipulated, then this application should be capable of creating a machine learning model capable of predicting which round prospects will be drafted in based on their combine performance.

### **Methodology**

This project will be developed using the waterfall methodology. This methodology will be the most cost-effective method for this project. The waterfall method is also appropriate for a project in which the requirements will be fixed throughout its development.

The requirements phase of the projects will include the initial discussions of features that will be included in the project. The design phase will include the outlining of how the features will be implemented and the data will be collected and used. This phase will include understanding how the data should be manipulated, and which machine learning model will be appropriate for the application. The implementation phase will include the actual manipulation of the data and the creation of the machine learning models. After completion of the implementation, the data manipulation, data visualization techniques, and the machine learning model creation will be tested for proper behavior. Finally, the project will pass into the maintenance phase where any remaining issues may be resolved.

### **Funding Requirements**

The cost to develop this model will be an upfront cost of $50,000, and a yearly fee of $100,000 for continual use. Maintenance and upgrades will be included throughout the duration of the contract between the Cleveland Browns and this company.

### **Stakeholders Impact**

This project will benefit the Cleveland Browns organization, the front office staff involved in making final decisions on draft day, and all members of the scouting department for the Cleveland Browns. The project will help any members in the scouting department in correctly evaluating the value of prospects based on their combine performance. The project will aid in the process of building the correct draft strategy for the front office staff. Overall, the project will help improve the Cleveland Browns ability to evaluate and strategically draft prospects, which will improve the Cleveland Browns organization.

### **Data Precautions**

The data used in this project will not contain any sensitive or protected data. Therefore, no existing frameworks for handling sensitive data will be applicable to this data. Some of the typical frameworks include HIPPA for handling health records, PCI DSS handling credit card information, and FERPA for handling educational records. These frameworks outline the ethical guidelines for handling data that is sensitive and must be protected. These guidelines include things such as role-based access, password protection, and various encryption standards.

### **Developer Expertise**

This project will be completed by a software engineer with a bachelor’s degree in computer science. The software engineer has also spent years self-studying data science and over ten years studying the process of evaluating NFL draft prospects. Along with the computer science degree, the software engineer possesses credentials of a certified project manager through an ITIL 4 Foundation Certification and a CompTIA Project+ Certification.

# **Prompt B**

## **Project Proposal**

**Problem Statement**

For NFL organizations to successfully build a strategy for selecting players at the NFL draft, it is imperative that the organization understands the value that each player holds come draft day. The NFL combine is the main testing process for NFL draft prospects. Current methods of evaluating value of prospects at the combine rely on reports from scouting experts. Using a historical analysis of how combine data translates to draft value can help to reveal the true draft value of each prospect.

**Customer Summary**

This application is meant to be used by the scouting department of the Cleveland Browns. The application is meant to supplement the traditional process of evaluating draft prospects. The application will require a very basic understanding of Jupyter Notebooks. Scouts for the Cleveland Browns will use the application to understand which round a player will be drafted based on historical combine and draft data.

**Existing System Analysis**

The current technology used by the Cleveland Browns scouting department is currently simple and basic. Scouts use word processing software, video editing software, and presentation software. Current technology is used to document areas of evaluation that scouts notice, view and compile video of prospects performing, and organize presentations of evaluations.

After completion of this project, each scout should be equipped with a Jupyter Notebook application capable of predicting where players will be drafted through machine learning. The application will use Python 3. All historical draft data and combine data will also be compiled into csv files and available to these scouts to use in conjunction with the Jupyter Notebook dashboard.

### **Data**

The data for this project will be collected from the NFL Play Statistics Dataset on Kaggle.com. The two csv files that will be collected from this dataset will be the combine.csv and draft.csv. There will be a few edits to these files necessary before they can be used. The draft.csv file will contain draft data from 1977 to 2019, but the combine.csv will contain data from 1987 to 2019. The only data that will be useful in this project will be draft data that corresponds to combine data, thus any draft data before 1987 will be removed from the file.

Both files will also contain certain unimportant data points that must be removed as well. The combine.csv file contains data columns that are irrelevant to actual data collected at the combine. These data columns will be removed. The only data columns necessary form the draft.csv file will be the draft round, draft pick, and combineId.

Once these files are properly formatted, they will be combined in the application. This combination will be executed using a left join of the combine.csv file and draft.csv file on the combineId of the prospects. This will ensure all undrafted players are included in the dataset, and that players who were drafted yet did not participate in the combine will be excluded.

This dataset that includes all the correct columns from each csv file will still have missing values throughout. Not every prospect will have a complete set of combine data or may have not been drafted in any round. Prospects drafted after round 7, or undrafted will be imputed with a value of 8 to give a numerical status of undrafted prospects. Missing age data will be imputed with the mean age of players at the draft. Any prospects that have a missing value for a 40-yard dash time will be removed entirely from the dataset due to the importance of this metric. College data will be encoded using binary encoding to allow it to be a useful categorical feature. The remaining missing values will undergo multivariate imputation. A Bayesian Ridge algorithm will be used to calculate a likely value for these metrics based on prospects with similar metrics in other areas.

### **Project Methodology**

The development of this project will follow the waterfall method. This method will allow a straightforward and cost-effective approach for a small project that will have its requirements remain constant throughout development. In following the waterfall method, this project will go through five phases of development which are requirements gathering, design, implementation, testing, and maintenance.

The requirements phase of the projects will include the initial discussions of features that will be included in the project. These discussions will include possible machine learning algorithms that may be useful and different ways that the data can be manipulated. The design phase will include the outlining of how the features will be implemented and the data will be collected and used. This phase will include actual decisions on data manipulation and a decision on which machine learning algorithm will be used. The implementation phase will include the actual manipulation of the data and the creation of the machine learning models. After completion of the implementation, the data manipulation, data visualization techniques, and the machine learning model creation will be tested for proper behavior. Verification testing will be done on each aspect of the application. Finally, the project will pass into the maintenance phase where any remaining issues may be resolved.

### **Project Outcomes**

The project outcome should be a fully functional dashboard in Jupyter Notebooks. This dashboard should clearly walkthrough the process of analyzing the relationship between combine data and NFL draft results. After all analyzations have been completed, the notebook should deliver models capable of predicting which round NFL prospects will be drafted in based on combine data.

### **Implementation Plan**

This project will be implemented from the bottom-up in a step-by-step fashion. First, the data will be collected and formatted correctly. Second, the data will be manipulated to fit the goals of the project. This will include both formatting, excluding, and imputing data. Third, various analyzations will be made and created. These analyzations will include descriptive methods of the data and meaningful graphical displays of the data. Fourth, the machine learning model creation process will begin. This step will be a cycle of training different models using different formats of data and a random forest classification algorithm. Each different model will be tested for accuracy and given a score. Finally, the models determined to be the most accurate will be finalized. The complete Jupyter Notebook walking through this process using Python 3, with the finalized models, will be the final deliverable of this project.

The application will require the installation of Jupyter Notebooks as well as various libraries that will be used throughout development. These libraries include scikit-learn, pandas, matplotlib, numpy, and category\_encoders.

### **Evaluation Plan**

Throughout the development process, different models will be created by using different combinations of combine data to find which combinations of data create models with the most predictability for which round prospects will be drafted in. The creation of these models will follow the rule of using 80% of the data for training the models and 20% of the data for testing the models. Each model will be given a score for how accurately it can guess which round each prospect will be drafted in based on the combine data it will be given. A score of around 70% accuracy will be used as the metric for a successfully created model. The models should be able to predict which round a player was drafted in with 70% or greater accuracy. The random forest classification model will continue to be trained with different formats of data until a score of around 70% or greater has been achieved.

### **Resources And Costs**

This project will not require many resource costs. The programming environment used will be Jupyter Notebooks ran on a 2020 MacBook Pro, which will not incur any extra costs. The application will be developed under the software developers own living conditions, which will also not incur any costs. The main costs will come from human resources. This project will require two months of development, and because of its valuable nature the development process will cost around $50,000.

### **Timeline And Milestones**

Development and deployment of this project is anticipated to take around 190 hours

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Milestone | Pre-requisi tes | Activity | Resource Assigned | Hours | Start | End |
| 1 | - | Requirement’s approval | Project Manager | 10 | 10/1/21 | 10/5/21 |
| 2 | 1 | Collection of Data | Software Engineer | 10 | 10/5/21 | 10/6/21 |
| 3 | 2 | Development Environment  Setup | Software Engineer | 10 | 10/7/21 | 10/8/21 |
| 4 | 2,3 | Data Manipulation | Software Engineer | 15 | 10/8/21 | 10/15/21 |
| 5 | 2,4 | Creation of Data Visualizations | Software Engineer | 25 | 10/16/21 | 10/28/21 |
| 6 | 2,4 | Final Data Preparation | Software Engineer | 10 | 10/29/21 | 10/30/21 |
| 7 | 6 | Machine Learning Model Creation | Software Engineer | 30 | 11/1/21 | 11/21/21 |
| 8 | 7 | Model Accuracy Testing | Software Engineer | 30 | 11/1/21 | 11/21/21 |
| 9 | 7,8 | Finalization of Model/Models | Software Engineer | 15 | 11/22/21 | 11/24/21 |
| 10 | 9 | Finalization of Application | Software Engineer | 15 | 11/25/21 | 11/27/21 |
| 11 | 10 | Full Application Test | Quality Assurance | 10 | 11/28/21 | 11/30/21 |
| 12 | 11 | Final project delivery | Project Manager | 10 | 12/1/21 | 12/2/21 |

# **Prompt C**

## **Application Files**

\NFLDraftProject

NFLCombineNotebook.ipynb Main data product dashboard

combine.csv CSV file of all combine prospect data from 1987 to 2019

draft.csv CSV file of all draft data from 1987 to 2019 including the round each player was drafted, the pick they were drafted, and the combine ID associated to the player

# **Prompt D**

## **Post-implementation Report**

### **Project purpose**

The NFL draft model project was created to help in the decision-making process that goes into an NFL draft strategy. It is imperative that NFL organizations understand the value of players in order to make the most efficient decisions on draft day. The NFL draft model project resulted in the creation of three different machine learning models that will predict the round a prospect will be drafted in based on past combine data for other prospects. One model can predict when running backs, wide receivers, safeties, and cornerbacks will be drafted, another can predict when linebackers will be drafted, and one more that can predict where edge players will be drafted. These models can be used by an NFL organization to help assess the value of players when building a draft strategy.

Graphical user interface, text, application, email

Description automatically generated

Figure 1: Skill Position Model Creation and Mean Accuracy Score

Graphical user interface, text, application, email

Description automatically generated

Figure 2: Linebacker Model Creation and Mean Accuracy Score

Graphical user interface, text, application, email

Description automatically generatedFigure 3: Edge Player Model Creation and Mean Accuracy Score

### **Datasets**

Graphical user interface, application, table, Excel

Description automatically generated

Figure 4: Combine Raw Data

Graphical user interface, application, Word

Description automatically generated

Figure 5: Draft Raw Data

Table

Description automatically generatedFigure 6: Final Data Combining Combine and Draft Data

The data used in this project was pulled from two csv files that contained raw data of combine results for NFL prospects and NFL draft results each from 1987 to 2019. The data was originally pulled from the NFL Draft Play Statistics Dataset on Kaggle.com. After collecting the data from Kaggle, the CSV files were cleaned so that only the data columns that pertained to the NFL combine or the NFL draft remained. The final CSV files are included in the zip file of the project. Figure 6 shows how the two CSV files were loaded and combined using the combineId to make one useful dataset.

Throughout the Jupyter Notebook provided, various manipulations were made to the data. These manipulations included removing data, various imputations of missing values, and the encoding of college data. Name data and pick data were removed from the dataset because name data has no use in determining where a player will be drafted, and pick data contains a direct correlation to round data, which would result in machine learning models that aren’t making useful predictions. Any prospect that had a missing value for a 40-yard dash time was removed from the dataset because of the importance of this performance metric. Prospects that have a missing value for the round they were drafted in had this missing value replaced with an 8 to give a classifiable value to undrafted players. Players with a missing age value were imputed with the mean age of players that were drafted. College data was encoded using a binary encoder. This was done so that college data could be converted from a categorical feature to something that the computer would be able to read in the creation of the machine learning model. The remaining imputation used multivariate imputation to obtain values for missing values. A Bayesian ridge algorithm was used to calculate the most likely value for those missing values based on the most predictive feature for that prospect. Multivariate imputation was used because it made the data more useful and more accurate in creating the machine learning models.

Each step described above is walked through in full detail in the Jupyter Notebook.

### **Data product code**

The code follows steps that walkthrough the manipulation of the data, the visualization of the data, and finally the creation of the machine learning models.

The manipulation successfully formats the data in a way that is most beneficial to the creation of the machine learning models. This process includes the combining of the CSV files as seen in Figure 6 and follows with all the further data manipulation decisions described in the Datasets section of this report.

The visualization portion of the code breaks down different relationships in the dataset using graphs and tables. This portion of code is used to understand how the relationships in the data lead to meaningful correlations between the combine data and the round prospects were drafted in. This portion of data was also used to identify reasoning to separate data by position groups to benefit the creation of the machine learning models.

The creation of the machine learning models is the last and most crucial section of the code. This section walks through various attempts in creating Random Forest Classification models that predict which round players will be drafted based on combine metrics. Random Forest Classification was decided to be the most beneficial model to use in this scenario. After deciding on the model, many test trials were carried out with different training sets. These test trials were executed to find which format of data would create the most predictive model. Data was separated and combined into different variations of position groupings, and different variations of removing data columns were attempted as well. The final decision was made to create three final models. One model for running backs, wide receivers, cornerbacks, and safeties, another model for linebackers, and another model for edge players. Training these models with and for these specific position groups was decided to be most accurate. Different data columns were further decided to be removed to improve the accuracy of these individual models. The Jupyter Notebook progressively walks through and describes this entire process.

### **Hypothesis verification**

This project was carried out under the hypothesis that If enough combine data and NFL draft data are correctly formatted and manipulated, then the dashboard created in this project should be capable of creating a machine learning model capable of predicting which round prospects will be drafted in based on their combine performance. The hypothesis was set to be evaluated based on a mean accuracy score of around 70% for the machine learning models. Based on this evaluation metric, each machine learning model in Figure 1, Figure 2, and Figure 3 proved the hypothesis to be true. Each model’s final verification score was above 80%.

### **Effective visualizations and reporting**

The Jupyter Notebook application provided effectively walks through each step of data preparation, manipulation, and visualization. The notebook begins by displaying how the two CSV files will be combined into a data frame, as seen in Figure 6. The dashboard continues by stepping through each process and attempt of manipulating the data to and displaying the results to the dashboard. There is also an entire section breaking down the data and analyzing it using various graphical elements. These graphical elements include histograms, line graphs, descriptive tables, and boxplots. Each visual element is used to draw conclusion about how different metrics will have different meaning based on the position group. 40-yard dash times were shown to have a clear correlation with which round players were drafted in, and this correlation was further shown to have a different meaning based on which position ran the 40-yard dash.

### **Accuracy analysis**

The accuracy of the three models created in this Jupyter Notebooks application are displayed in Figure 1, Figure 2, and Figure 3. A final model creation score is provided at the bottom of these images, but the main accuracy score for these models was assessed based on an algorithm that built 100 different models using 100 different variations of data. This algorithm separated each variation of data into a split of 80% training data and 20% testing data. The algorithm then found a mean accuracy for all 100 models based on the test sets of data and returned the average score of all 100 models. This code is commented out in the notebook because it takes a long time to process, and it was only necessary for evaluating a truer accuracy score for each model.

The truer accuracy score that these models produced was included as markdown in the notebook. The skill position model earned a score of .844, the linebacker model earned a score of .815, and the edge player model earned a score of .867. Before carrying out this project, a score of 70% was set to be the marker for a successful predictive rate. Each model successfully exceeds this rate.

### **Application testing**

Each portion of the model went through a final verification and integration test at the end to confirm the proper functionality of each portion of the dashboard. The data was ensured to be accepted and properly loaded into the dashboard during the final phases of testing.

The application was built chronologically step by step, so verification and integration testing persisted throughout each phase of development. Model testing ensued as a cyclical process towards the end of the development but was dependent on previous phases of testing.

# **Appendices**

## **Installation Guide**

1. Install the following prerequisite applications:

* Jupyter Notebooks
* Python 3 with supporting libraries: scikit-learn, matplotlib, pandas, numpy, and, category\_encoders

1. Extract NFLDraftProject.zip into the desired directory.
2. Open the Jupyter Notebooks application
3. Navigate to where NFLDraftProject.zip was extracted through Jupyter Notebooks.
4. Ensure that NFLCombineNotebook.ipynb, combine.csv, and draft.csv all reside in the same folder. This should be true if NFLDraftProject.zip was extracted and unmodified.
5. Open NFLCombineNotebook.ipynb

## **User Guide**

Graphical user interface, text, application, email

Description automatically generated

Figure 7: Final Dashboard Display

In order to run the project, first the installation steps described in the installation guide must be executed. After these steps are completed, the application can either be run step by step or ran all at once.

To run each portion step by step, start at the very first cell and continuously press the run button in the top left corner of the page. It is imperative that each cell is ran in order to ensure proper functionality.

To run each cell all at once, navigate to Cell tab in the top left corner and click on the Run All option in the drop-down menu. Running all cells at once will ensure that each portion of code is executed chronologically as intended.

## **Summation of Learning Experience**

The scope of this project was extremely intimidating at first, but I was able to use prior knowledge in combination with the exploration of new material to see the project through. My prior skills in software development and python came in handy throughout this process. However, the main bulk of learning that went underway came from self-study on machine learning. I explored different areas of machine learning using FreeCodeCamp, Udemy, and LinkedIn Learning. Using these resources, I gained insight in how to build a data application using Jupyter Notebooks.

The rest of my learning was carried out through the actual development in the notebook. I was introduced to new ways to use python libraries in order to display and manipulate data. I also underwent many hours of testing different machine learning algorithms and how the behavior of these algorithms would change based on the format of data the algorithms were trained on.

Overall, this was the most intense project I completed throughout my degree program. The knowledge I gained was more extensive than I expected, and I trust that it will only benefit me moving forward.