College Football 4th Down Prediction Report

Introduction:

In college football, fourth down decisions can be crucial to winning a tightly contested game. Coaches must decide whether to go for a risky first down or take a safer alternative. Previous studies like Romer [1] used NFL data to create a dynamic programming approach that assessed and scored different fourth down strategies based on field positioning. Lock and Nettleton [2] expanded on this by using random forests to estimate win probability before each play in an NFL game by incorporating situational variables. Jennifer Wright [3] looked at average punting distance and how it influenced whether to punt or not on fourth down. However, all three studies focus on the NFL and do not account for the nuances of college football; immediate applicability for fourth-down-conversion decisions is still limited. This highlights the need for predictive models, specifically tailored to college football, that assist coaches with real-time, fourth-down recommendations.

Dataset:

We examine play-by-play data from the College Football Data API, incorporating features such as field position, time remaining, score differential, and distance to the first down to evaluate whether a 4th down setting is favorable for a conversion.

Problem Definition:

The goal of this project is to predict whether the Georgia Tech football team should "go for it" on fourth down by building a binary classification model. Using play-by-play data and situational factors, the model will provide data-driven recommendations during critical fourth-down plays.

Preprocessing Methods:

The preprocessing methods we plan to use are data cleaning, feature engineering, and imbalance data. Data cleaning will ensure the dataset is free from errors (missing values, inconsistencies, etc.). Feature engineering will help the model make more accurate predictions because the raw data might not capture the complexities of decision making. Imbalance data will prevent the model from being biased toward the majority class. Since 4th down attempts happen much less than punts/field goals, the model might be biased if we don’t address this.

ML Algorithms:

The first ML Algorithm we plan on using is KMeans. This can be used to cluster similar game situations together to analyze common patterns on 4th down decisions. We’ll also test a random forest classifier, following the findings of the Amazon ML Solutions team [4] in using a decision tree-based model. To avoid the additional requirement of scoring the skill of each team per play that would be needed for a regression model, we select a support vector machine (SVM) as our third classifying model. We expect favorable conversion situations to be separable from those where a 4th down is likely unattainable by a hyperplane, as those “go for it” situations will have many similar values in the data.

Data Sourcing and Cleaning

For the collection of our data we retrieved Georgia Tech fourth down plays data from the College Football Data API. We decided to collect 15 years worth of information to make sure we had a robust enough dataset. For each year we retrieved the game ID, time remaining, field position, score differential, and distance to the first down for each play. We also implemented a classification mapping for different play types like punts or field goals which helped keep a cleaner and more refined dataset. By isolating only the fourth-down plays, we addressed the data imbalance challenge and started setting up our dataset for more targeted preprocessing and feature engineering.

Potential Results and Discussion:

To evaluate our model's performance in predicting whether to convert on 4th down, we will use several metrics including accuracy, which will measure the overall correctness of the model, with a goal of achieving 80%. We'll also focus on precision to ensure the model minimizes false positives, aiming for over 75%. Recall is important to capture true positives, with a target of 70%. The F1 score will help balance precision and recall, aiming for 0.75 or above, while ROC AUC will provide a general measure of model performance, with a goal of 0.85 or higher. We expect the model to slightly favor conservative decisions but can be fine-tuned to balance risk and reward in critical moments. Finally, we could use a confusion matrix to help us visually see where our model is getting things right or wrong, since this matrix breaks down the model's performance into true and false classifications. We believe that reaching these thresholds would constitute an accurate model for converting on 4th downs.

Results and Discussion:
K-means clustering:

Even though the end goal of this project is to create a binary classification model of whether to "go for it" or not on fourth down, we decided to use K-means clustering with a value of k greater than two in order to capture more of the complexity in the data. By clustering the samples into more than two groups, our goal was to explore different levels of risk and context that will ultimately help us improve the accuracy and reliability of the final decision making model for fourth down plays. These k-means classifications can provide additional insights into our data and could potentially serve as features in the final model to help it learn more complex patterns and make better decisions.

In order to determine the optimal number of clusters, we began by looking at the inertia for different values of k burning the elbow method. We plotted the inertia values for k ranging from one to 30 as seen in **Fig. 1**.

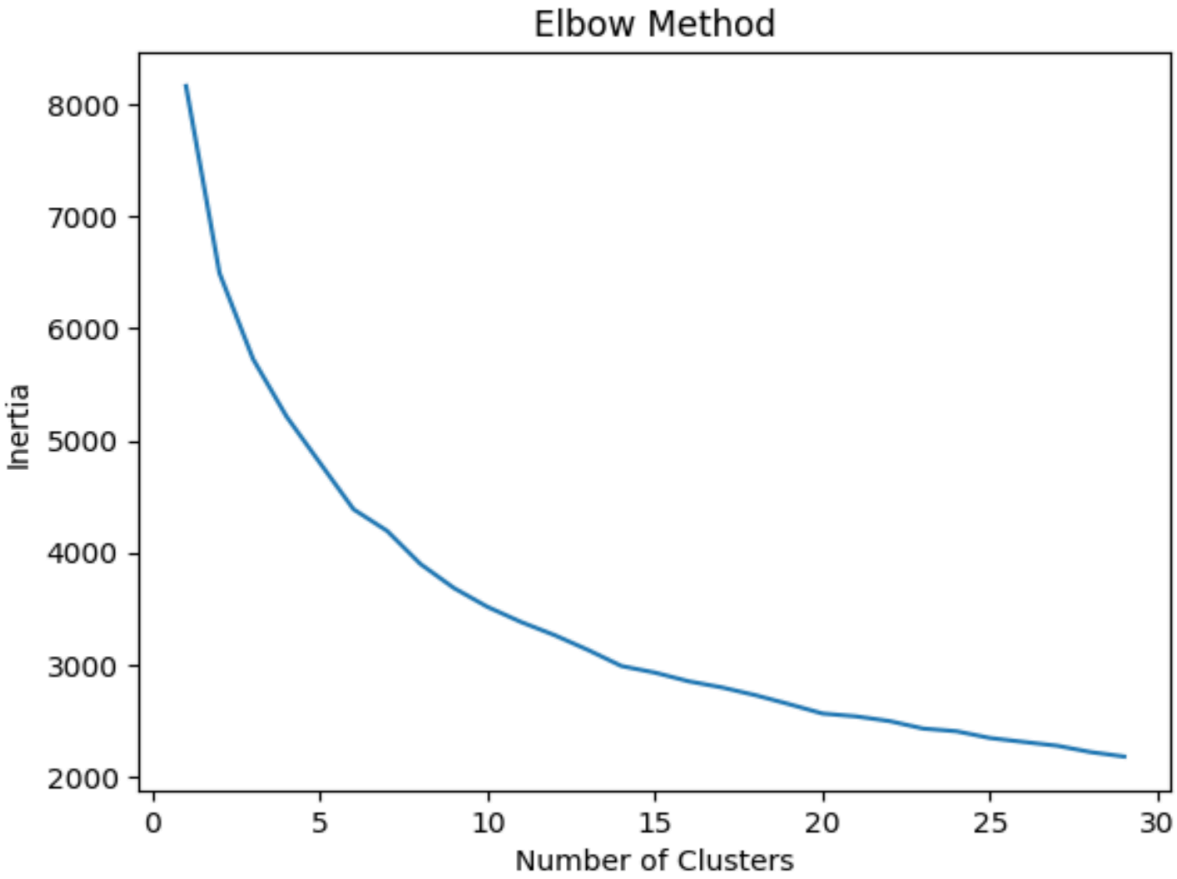

Fig. 1. Elbow curve for determining the optimal number of clusters.

Based on **Fig. 1** above, it appeared that the ideal number of clusters was somewhere between four and seven clusters. The curve started to flatten out after this range, which means that more than 7 clusters would not substantially improve the separation of our data.

To decide specifically on how many clusters to pursue, we evaluated three different clustering metrics for k = 4, 5, 6 and 7. First, we looked at Silhouette Score which is a measure of how similar each point is to its own cluster compared to other clusters. Next, we examined the Davies-Bouldin Index to gain a better understanding of cluster compactness and separation. Lastly, we calculated the Calinski-Harabasz Index to look at the variance ratio between clusters. Results for Silhouette Scores, Davies-Bouldin indices, and Calinski-Harabasz indices are shown in **Fig. 2**, **Fig. 3**, and **Fig. 4**, respectively.

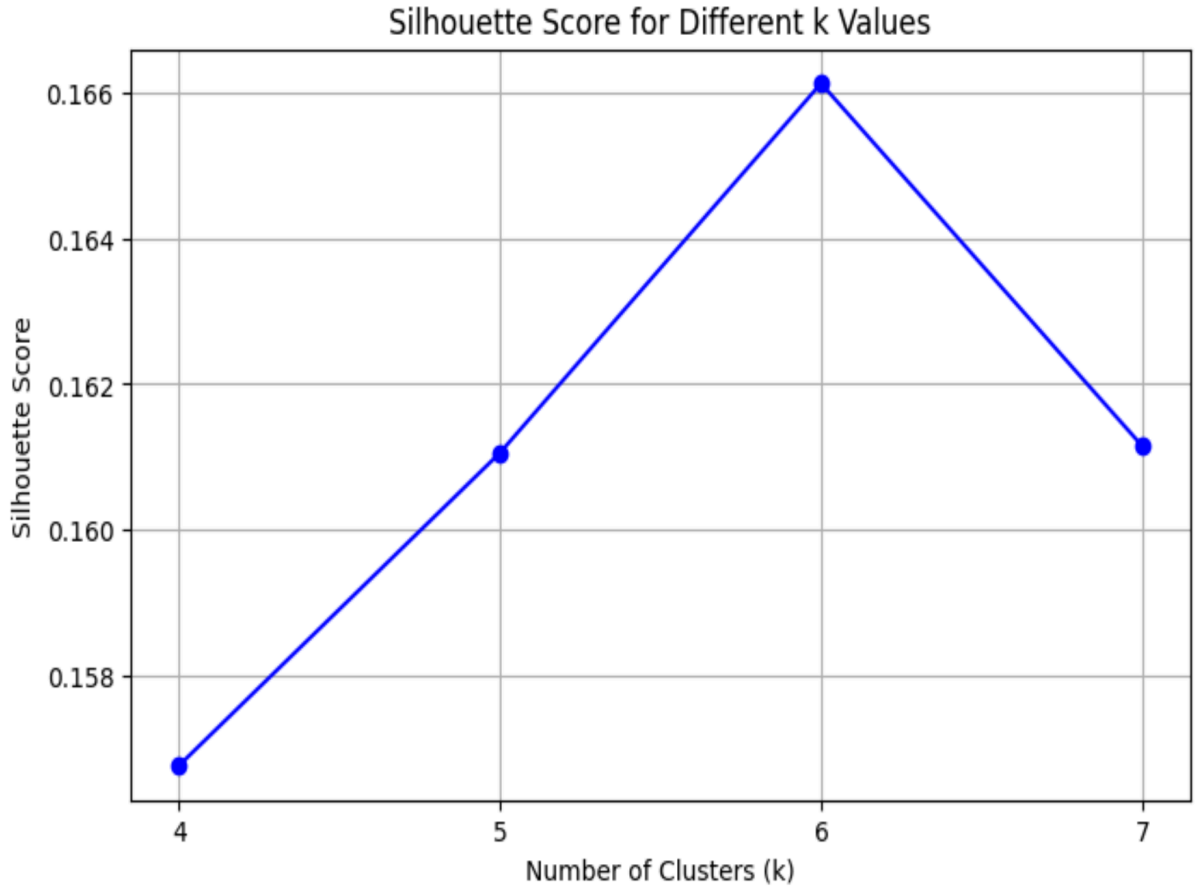

Fig. 2. Silhouette Score for different values of k. The score is highest at k = 6 which indicates better cluster separation.

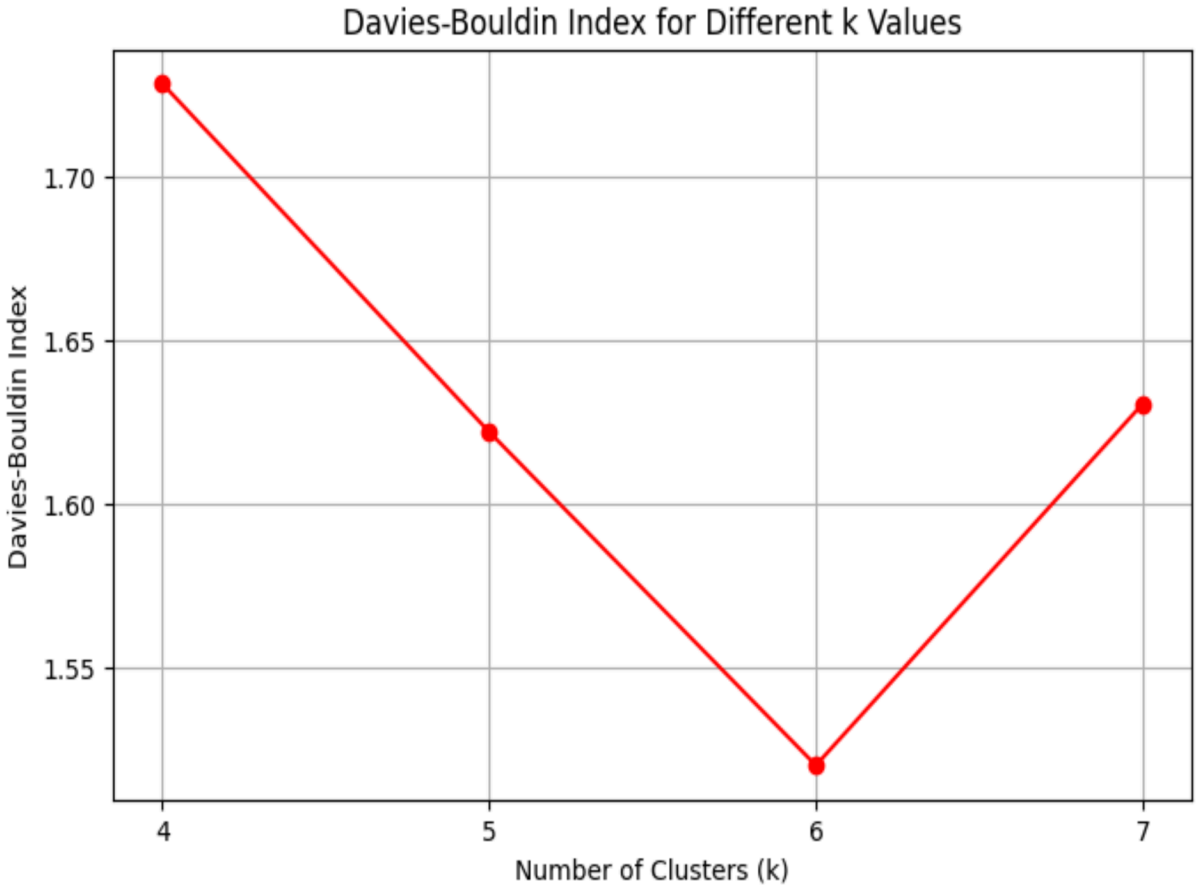

Fig. 3. Davies-Bouldin Index for different values of k. The lowest value is seen at k = 6 which suggests the clusters are most compact and well-separated at this point.

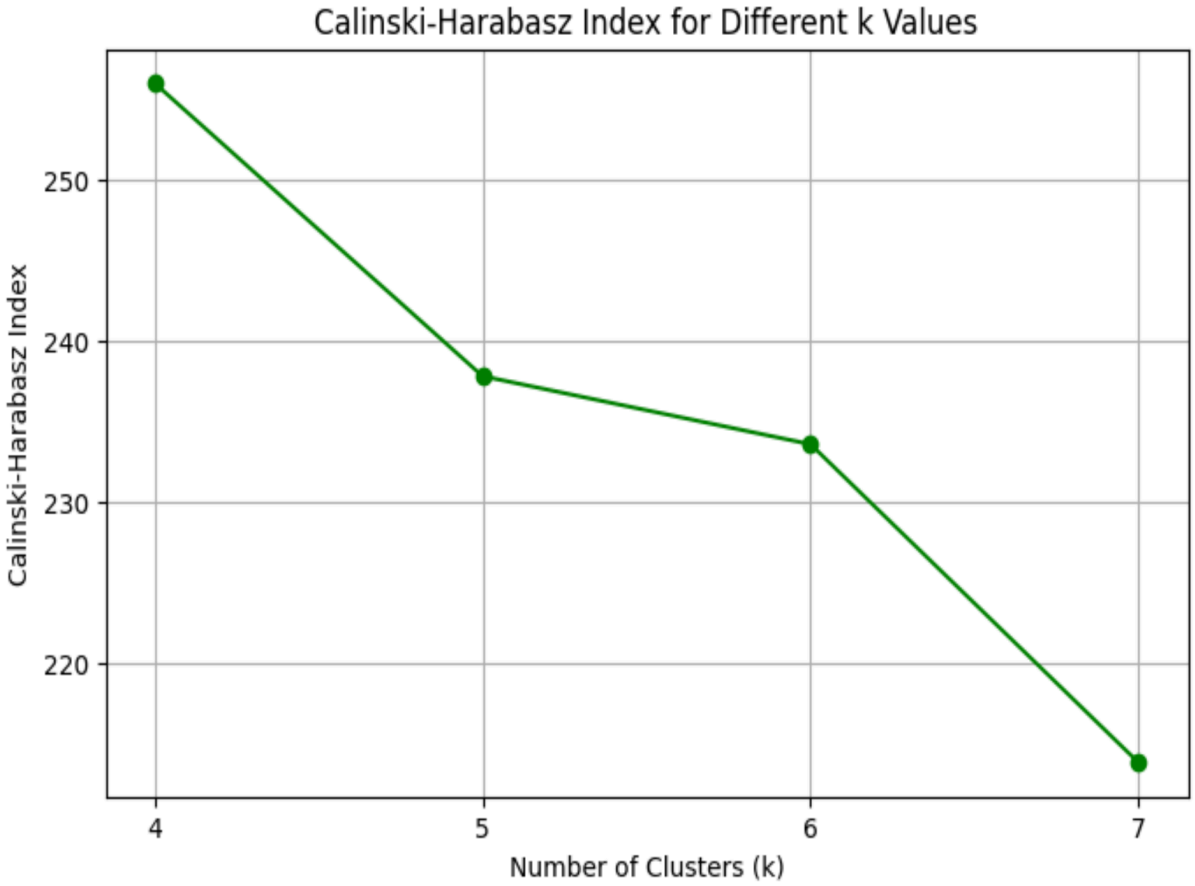

Fig. 4. Calinski-Harabasz Index for different values of k.

The silhouette score was highest at k = 6 which suggests that the separation between clusters was best when k was set to 6. The Davies-Bouldin Index was also lowest at k = 6 which means clusters were more compact and distinct from one another compared to other values of k. Lastly, the Calinski-Harabasz Index was highest at k = 4, but the difference was not significant enough to override the improvements in Silhouette Score and Davies-Bouldin that we saw at k = 6. Based on these metrics and the graphs shown above, we chose k = 6 as our optimal number of clusters because it provided the best balance between compactness and separation of clusters.

After deciding on k = 6, we ran the K-mean clustering algorithm and used Principal Component Analysis (PCA) to reduce our data down to two dimensions, solely for visualization purposes. This PCA projection allowed us to visualize the cluster in a 2D space to better understand how the data points were grouped into six different clusters. The scatter plot shown below(Figure 5) definitely showed some overlap, but overall, it gave us relatively well-defined clusters that capture the variability in the data.

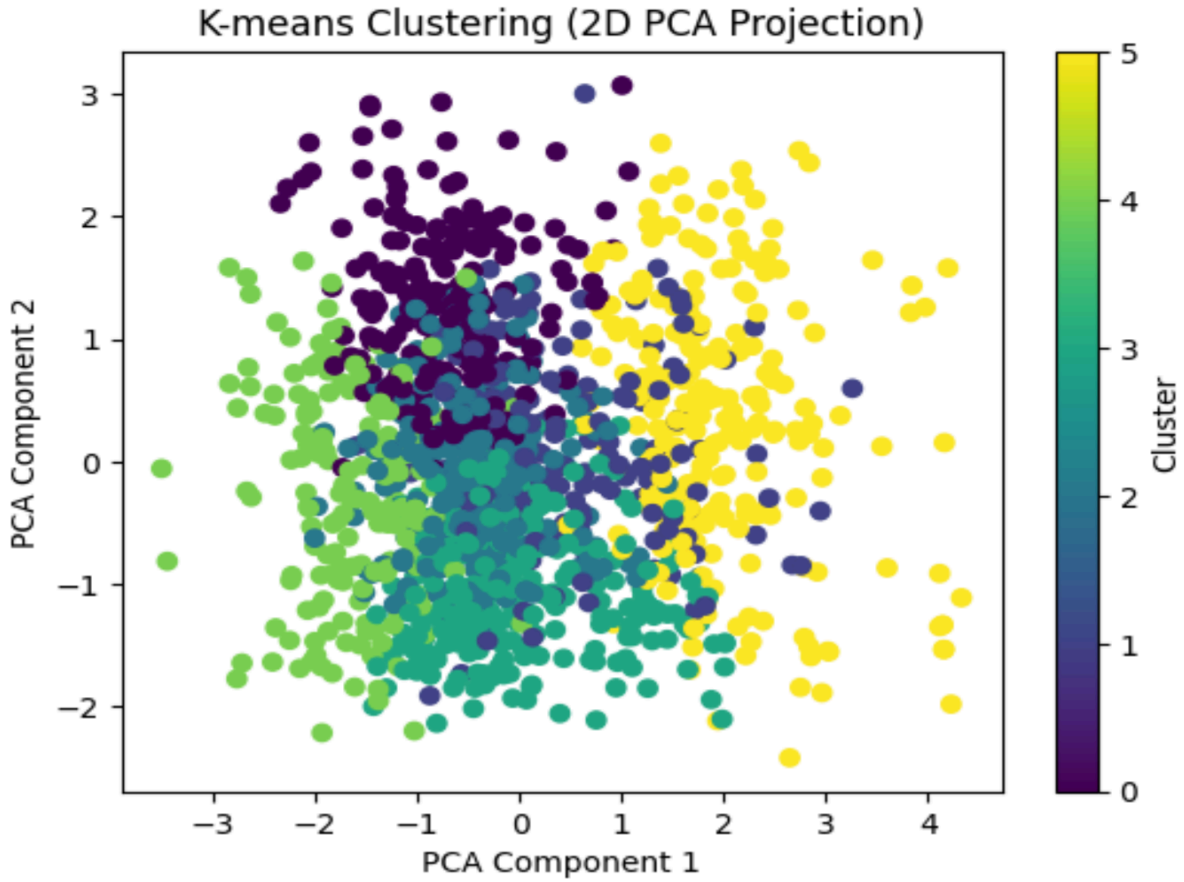

Fig. 5.PCA projection of the K-means clustering with k = 6.

For k = 6, our clustering had a silhouette score of 0.166 indicating the clusters had a moderate level of separation, a Davies-Bouldin Index of 1.52 suggesting the clusters were well separated and compact, and a Calinski-Harabasz Index of 233.58 which tells us the between-clusters variance was captured well by our model.

In the end, using K-means clustering with k = 6 allowed us to use k-means as more than a simple binary classification but rather a tool to better understand the various levels of risk and opportunity in the data. We believe this approach has provided us with better insights into our highly complex dataset which will be incredibly beneficial in making data-driven decision in these critical gametime situations.

Random Forest:

The second model we used was the random forest classifier. This classifier was used to predict whether Georgia Tech should "go for it" or "not go for it" on fourth down. Random forest is a model that builds multiple decision trees for more robust classification. This is particularly useful when handling complex interactions between features such as score, field position, and time. Additionally, random forest is an ideal model for binary classification such as this one.

Two approaches were evaluated: one using the original dataset and another that incorporated the K-means clusters as an additional feature. The performance of both models were assessed using various quantitative metrics such as accuracy, precision, recall, and F1-score, and confusion matrices were generated to give a visualization of the model's performance.

Overall, the metrics indicate that the random forest classifier was effective in determining whether Georgia Tech should go for it or not on fourth down. Without the K-means cluster as an additional feature, the random forest achieved an **accuracy** of 87.91%, **precision** of 87.39%, **recall** of 87.91%, and an **F1-Score** of 87.25%. This indicates that the model had a solid performance of classifying plays correctly. The confusion matrix generated from this model is shown below.

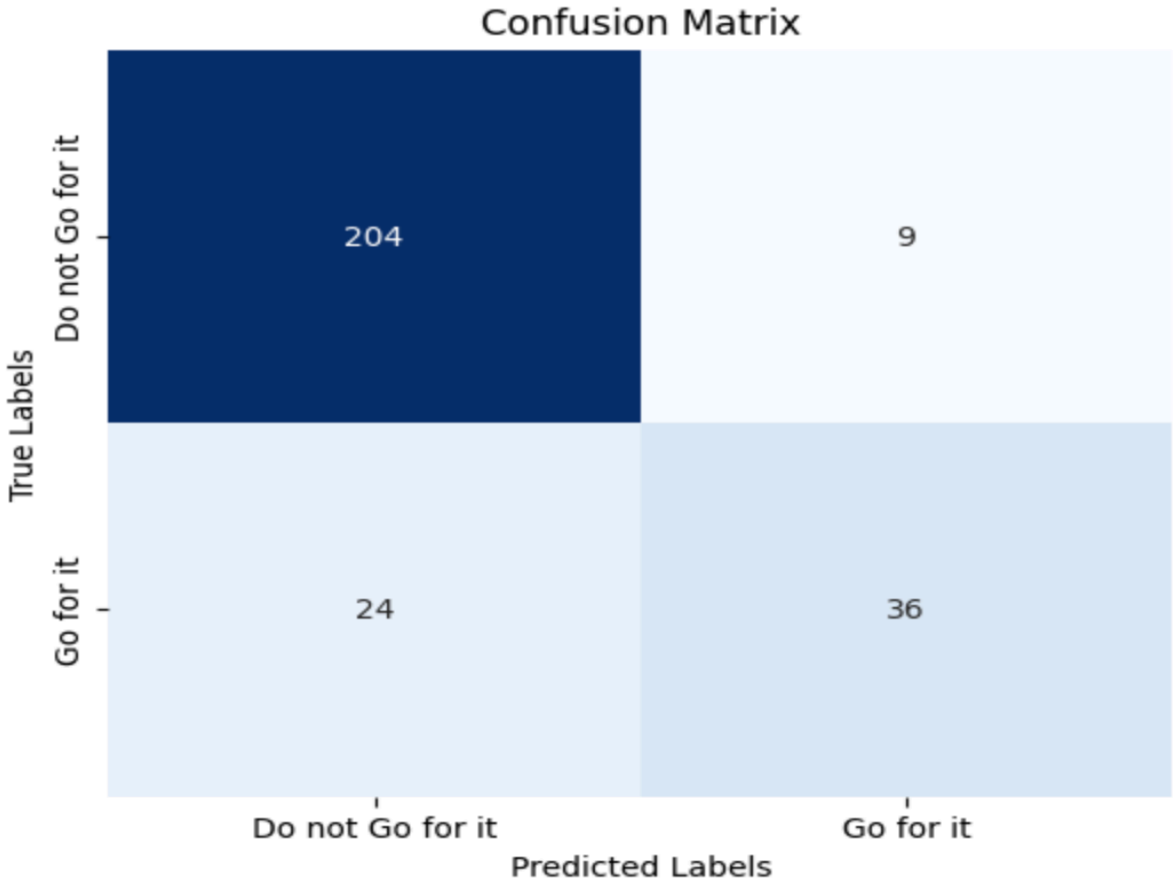

Fig. 6. Random Forest Confusion Matrix without K-means clusters.

Based on the confusion matrix, the model is able to accurately predict true positives and true negatives for the most part. The number of false positives and false negatives are small compared to the number of true positives and true negatives.

Although this model does a good job, it can do even better. We trained another random forest model, this time with the clustered data from K-means as an additional feature in the dataset for enhanced feature representation. This model performed much better than the first model achieving an **accuracy** of 97.44%, **precision** of 97.52%, **recall** of 97.44%, and an **F1-Score** of 97.38%. This model clearly performed much better based on these metrics compared to the first model. The confusion matrix generated by this model also shows less false positives and false negatives as shown below.

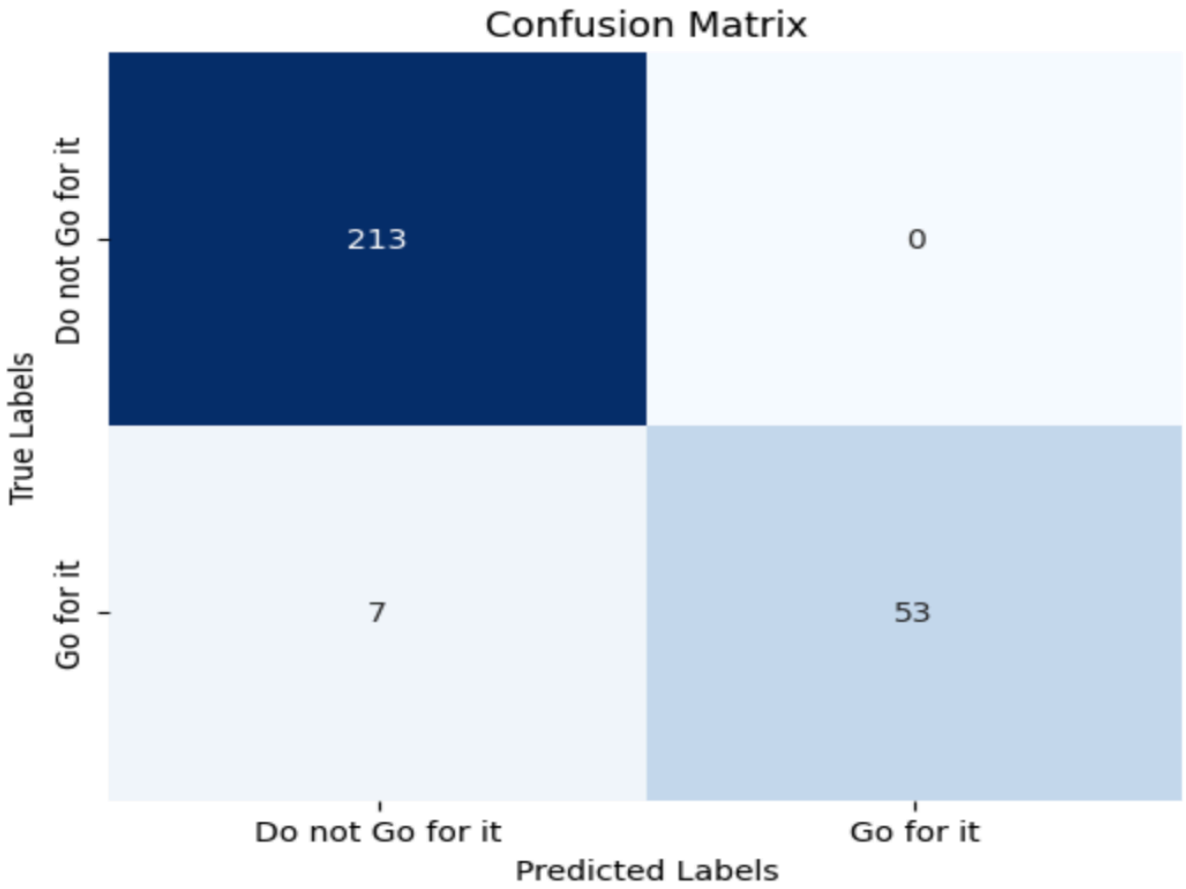

Fig. 7. Random Forest Confusion Matrix with K-means clusters.

This confusion matrix shows that there were 0 false positives and only 7 false negatives. Therefore, the addition of the K-means clusters allows the model to better differentiate between the two outcomes. There are a few trade offs though. For example, the first model that trained on the original data is simpler, faster to train, and is less computationally expensive. In contrast, the second model that uses the K-means clusters is more computationally expensive and requires additional preprocessing. Additionally, the addition of clusters could provide a risk for overfitting if the clusters were not meaningful. Moving forward, further experimentation with other feature engineering techniques or ensemble methods could provide additional insights.

Support Vector Machine (SVM):

The Support Vector Machine (SVM) model was trained with a focus on play success, which demonstrated its effectiveness at handling complex datasets and capturing nuanced patterns to distinguish between risky and conservative plays. Through hyperparameter tuning (e.g., C = 0.1, balanced class weight, gamma = 0.1, and a polynomial kernel), the SVM achieved an accuracy of 83% and a macro F1-score of 0.71. These metrics indicate a decent performance but reveal challenges due to class imbalance, particularly in predicting risky "go for it" plays. The model excelled in identifying conservative punts, with high precision (0.88) and recall (0.92), but struggled with risky decisions, achieving lower precision (0.60) and recall (0.47). A key limitation was the lack of robust data on punt outcomes, introducing uncertainty in evaluating its predictions. These results emphasize the need for further balancing techniques and expanded data collection to improve predictions and decision-making in critical game situations.

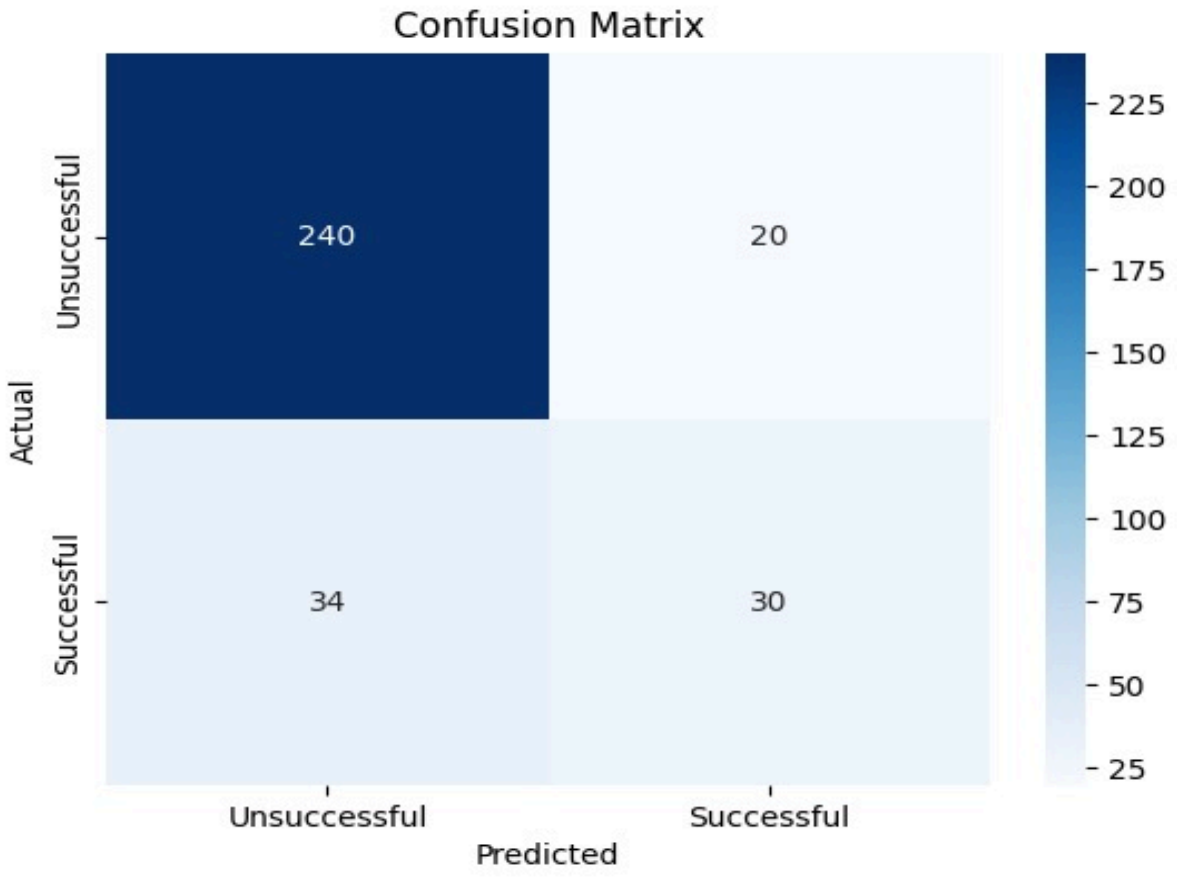

Fig. 8.SVM Confusion Matrix.

The confusion matrix illustrates the SVM model's performance in predicting fourth-down plays, highlighting its strength in identifying unsuccessful plays and its challenges in predicting successful ones. The model achieved 240 true negatives and 30 true positives, demonstrating its ability to classify conservative decisions accurately. However, it also produced 34 false negatives, where successful plays were missed, and 20 false positives, where unsuccessful plays were incorrectly predicted as successful. These results reflect the class imbalance in the dataset and the difficulty the model faced in capturing risky "go for it" scenarios effectively. This visualization underscores the need for additional data and improved techniques to enhance prediction accuracy for critical game-time decisions.

Conclusion:

In this project, we successfully developed machine learning models tailored to address the nuances of college football fourth down decisions. Among the three models we tested, the Random Forest classifier trained on pre-clustered data (using our k-means clustering) demonstrated the best performance after achieving a 97% accuracy. On the flip side, the SVM model highlighted the challenge of predicting riskier plays– specifically the “go for it” scenarios– because of class imbalances. Going forward, incorporating additional features, like punt outcomes, and integrating more formal techniques to counter any class imbalances will be incredibly beneficial to further improving the performance of these models. Overall, these models provide a solid foundation for data-driven recommendations that will help our Georgia Tech football coaches make smart decisions during critical moments and beat UGA next year.

References:

[1] D. Romer, "It's Fourth Down and What Does the Bellman Equation Say?: A Dynamic-Programming Analysis of Football Strategy," National Bureau of Economic Research, Jun. 2002.

[2] D. Lock and D. Nettleton, "Using random forests to estimate win probability before each play of an NFL game," Journal of Quantitative Analysis in Sports, vol. 10, no. 2, pp. 197–206, 2014.

[3] J. A. Wright, "Fourth down decisions in NFL football: a statistical analysis," M.S. thesis, California State University, Northridge, Aug. 2007.

[4] S. Senthivel et al., "Next Gen Stats Decision Guide: Predicting fourth-down conversion," AWS Machine Learning, Nov. 18, 2021. [Online]. Available: https://aws.amazon.com/blogs/machine-learning/next-gen-stats-decision-guide-predicting-fourth-down-conversion/. [Accessed Oct. 1, 2024].

Link to Gantt chart: Gantt.xlsx

Group Members and Contributions:

Group Member	**Contribution**
George Corbin	Data sourcing, data cleaning report
Shreya Khurjekar	Clustering metrics, visualizations, K-Means report
William Tjokroamidjojo	K-Means clustering model & plotting
Scott Williams	Data cleaning, SVM selection report & HTML translation
Truman Yardley	Data sourcing & filtering