The dataset under scrutiny is expansive, containing a myriad of variables, but the primary target variable of interest is 'expected points added.' To navigate this vast array of information, 35 distinct features were available, each describing various facets of a play. To streamline the analysis, a correlation plot was employed to identify numerical features that exhibited correlations with the target variable. From this analysis, specific features were singled out: 'down,' 'passLength,' 'penaltyYards,' 'prePenaltyPlayResult,' 'playResult,' and 'passProbability,' showcasing significant correlations as we can observe in figure 1.

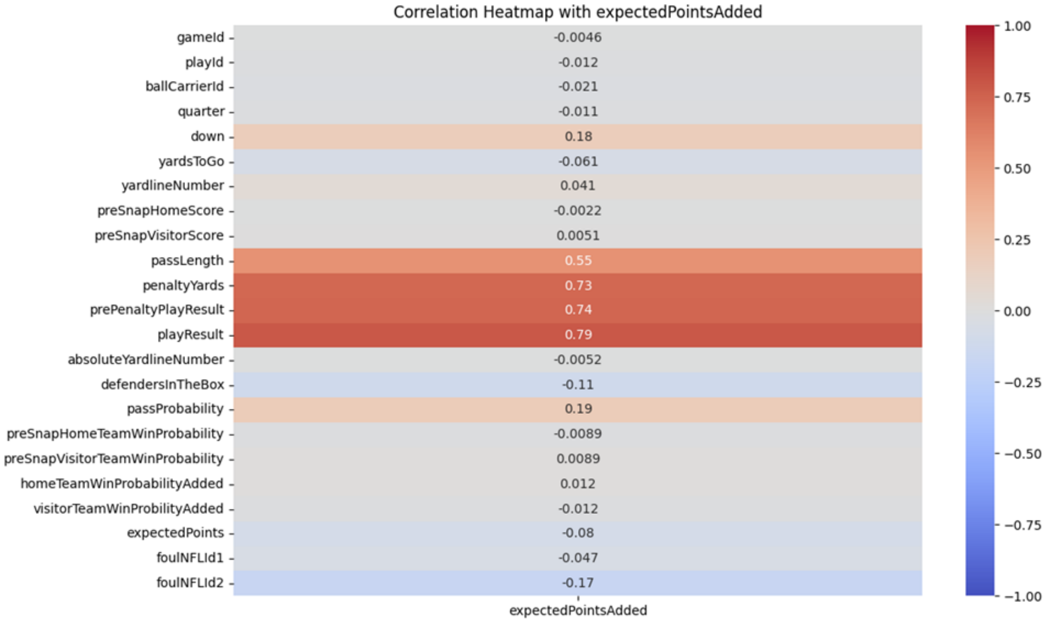

Figure 1: Numerical Correlation Plot

However, the dataset wasn't solely composed of numerical values; categorical data were present and required encoding for correlation assessments. Among the encoded categorical data, 'passResultEncoded' and 'isTouchdown' emerged as a pivotal featurs, as we can see in figure 2. The 'passResult' column, initially indicating pass completion, incompleteness, or handoff, underwent encoding to assign values: completed (1), handoff (0), and incomplete (-1). Concurrently, 'isTouchdown' was ingeniously engineered by parsing through the 'playDescription' column, discerning whether a play culminated in a touchdown or not. This ingenious engineering of features enabled a nuanced understanding of play outcomes and their association with varied factors.

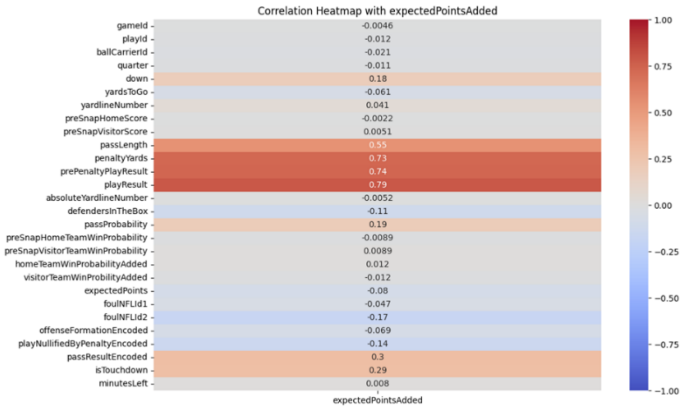

Figure 2: Correlation plot with encoded features

Furthermore, to distill the intricate relationships embedded within the dataset, dimensionality reduction techniques like PCA and UMAP were applied. Principal Component Analysis facilitated the reduction of feature dimensions while retaining key information, allowing for a more streamlined analysis without compromising the dataset's integrity. In figure 3 we see the scree plot of the PCA analysis, as we see at 8 components, we reach the maximum variance.

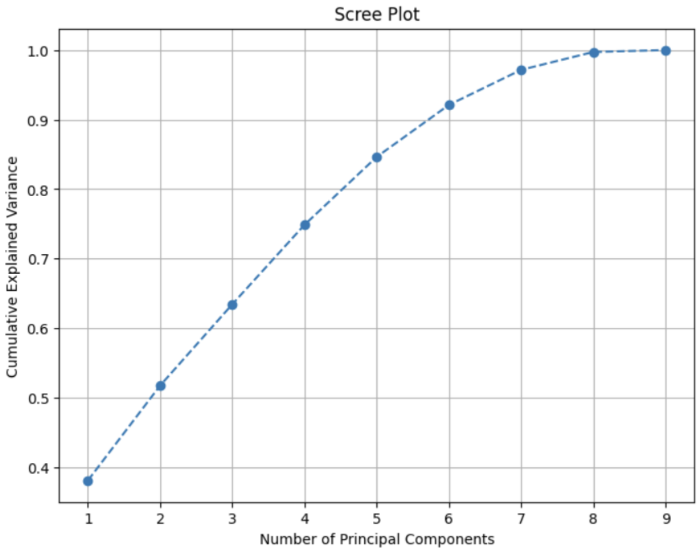

Figure 3: PCA Scree plot

Similarly, Uniform Manifold Approximation and Projection provided a low-dimensional representation of the dataset, preserving intrinsic structures and unveiling patterns that might have been concealed within the multidimensional data. As we see in figure 4, the UMAP shows that there is not really a very good separation of the data. This means that we will have to apply more complex models to properly describe the data.

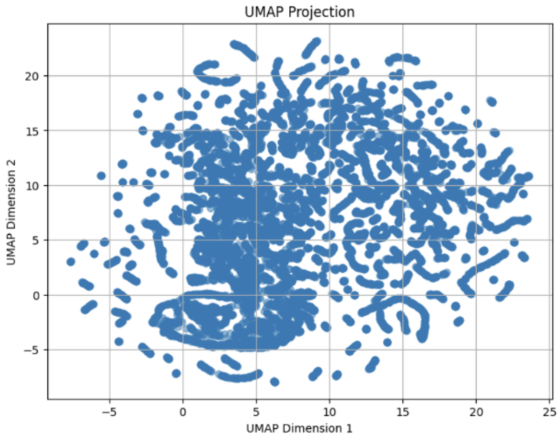

Figure 4: UMAP

We used six different regression models to predict the outcome of interest.

- Linear Regression Models: Traditional statistical models suitable for capturing linear relationships between predictors and the outcome.
- Polynomial Regression: An extension of linear regression where relationships between the independent variables and the dependent variable are modeled as an nth degree polynomial.
- Decision Tree Regressor: A model that uses a tree-like graph of decisions and their possible consequences.
- Neural Networks: A model that develops layers of neurons with different weights to approximate any type of function

- Support Vector Regression: A model that attempts to draw a hyperplane that fits the data while also drawing decision boundaries to encompass as much of the data as possible for prediction.
- Random Forest: A model that uses the output of many decision trees and consolidates it to one output.

In our quest to model and understand the dataset better, we opted for various regression techniques, each offering unique advantages tailored to different aspects of the data. Linear regression became our primary choice due to its proficiency in predicting continuous variables, aligning seamlessly with our target variable's nature, which demanded such predictions. The selection of Polynomial Regression followed suit, as it accommodates data that deviates from linear patterns, allowing for a more flexible fit to our diverse dataset.

Furthermore, Decision Tree Regressor was an integral choice owing to its inherent interpretability, facilitating a clear understanding of the model's decision-making process. Given the complexity embedded within our dataset, the implementation of a Neural Network was imperative. Its capacity to navigate intricate data patterns and adapt to complex structures perfectly aligned with our dataset's demands.

Exploring further, we experimented with Support Vector Machines (SVM) to discern potential separations within higher dimensions of the data. Additionally, we applied Random Forest, leveraging its bagging technique to assess its efficacy in handling our dataset's complexity. Each model choice was a deliberate attempt to gain deeper insights and extract the most comprehensive understanding of the dataset's underlying patterns and relationships.

Model Evaluation

Mean Squared Error (MSE) is a common metric for evaluating regression models, measuring the average squared difference between the estimated values and the actual value. We calculated the MSE for each of the six models. While the exact values and comparison among models are not evident in the first few cells, the conclusion that the MSE for all models was not satisfactory suggests that the current predictors in the dataset may not be sufficiently capturing the complexities of the outcome.

Importance and Implications

The use of diverse models in this project is significant. Each model comes with its assumptions and strengths. For instance, linear and polynomial regressions are interpretable and good for understanding linear relationships, while decision trees can capture non-linear patterns without needing feature scaling. Neural networks, although less interpretable, are powerful in capturing complex patterns in large datasets.

However, the unsatisfactory MSE across all models suggests a need for further exploration. This could involve Feature Engineering by creating new features or transforming existing ones to better capture the underlying relationships, exploring more sophisticated models or fine-tuning the existing models' hyperparameters, and ensuring the data is of high quality and considering the addition of more data points for better training.

The project is a comprehensive attempt to apply machine learning to a complex real-world problem in the context of NFL plays. The choice of diverse regression models reflects a thorough approach to tackle the prediction task. However, the overall unsatisfactory MSE highlights the challenges in predicting outcomes in sports analytics, where outcomes are influenced by many nuanced and interdependent factors. The project paves the way for further exploration, both in terms of data and modeling techniques, to enhance the predictive accuracy and gain deeper insights into the factors contributing to a team's likelihood of scoring.