**Introduction & Dataset Acquisition:**

Good evening, everyone. Today, we're excited to present our work on building a machine learning algorithm to predict employee attrition and departure timeframe.

We started by collecting relevant data from IBM Human Resources records. This dataset contains various attributes of employees, such as age, job satisfaction, salary, and performance ratings, as well as a lot of other information.

Our goal is to assist decision-making in the HR department by providing insights into when to initiate the hiring process and strategies to retain valuable employees.

**Data Preprocessing & Exploration**

After acquiring the dataset, we performed several preprocessing steps, including handling missing values, encoding categorical variables, and scaling numerical features. Additionally, we mapped the targeted variables, to binary values 0,1 to facilitate model training.

Before diving into our model, let's take a look at the features we selected for our analysis.

. We conducted visualizations using Tableau to explore various employee attributes and their correlations with attrition and departure timeframe.

. All features selected to be applied to the algorithm had a high impact on our target value. And we confirmed this by testing multiple features collections until finding the best accuracy.

. These features were chosen based on their potential impact on employee attrition, as identified through our exploratory analysis.

**Using python, we created a code to take any random 100 rows from our original dataset and drop the attrition column so we can use this dataset for testing and applying our algorithm and so we can check the accuracy of the outcome.**

**Employee attrition Initial Approach was Neural Network**

We initially employed a neural network model for attrition prediction.

However, we encountered challenges such as overfitting and limited performance metrics and evaluations.

**Transition to Random Forest Model**

To address these challenges, we transitioned to a random forest classifier for attrition prediction.

The random forest model offered better performance, simplicity, and interpretability compared to the neural network.

**Key Changes**

We summarized the changes made in transitioning from a neural network to a random forest model:

. Model selection change.

. Data balancing with RandomOverSampler.

. Feature engineering.

. Hyperparameter optimization using grid search.

. Interpretability and performance metrics.

(The goal was to improve the model's performance metrics, such as accuracy, precision, recall, or F1-score, by finding the optimal values for hyperparameters like learning rate, regularization strength, number of hidden layers, and so on)

Data Reading and Preprocessing:

* We read the dataset.
* mapped to numeric labels (0 for 'No', 1 for 'Yes').
* Features relevant for attrition prediction are selected and stored in the variable X.
* The target variable for attrition prediction, 'Attrition', is stored in the variable Y.

Data Splitting:

* dataset is split into training and testing sets using the (train\_test\_split) function from scikit-learn.
* The split is done with a test size of 20% and a random state for reproducibility.

Data Balancing:

* The class distribution is balanced using the RandomOverSampler technique from the imbalanced-learn library. This is done to address any class imbalance issues. ( only 19% of the data is Attrition yes)

Data Scaling:

* Standard scaling is applied to the training and testing features using StandardScaler from scikit-learn.
* Scaling ensures that all features have a similar scale, which helps improve the performance of the Random Forest Classifier.

Hyperparameter Optimization:

* A parameter grid containing different hyperparameters for the Random Forest Classifier is defined.
* RandomizedSearchCV is used to perform hyperparameter optimization. It searches for the best combination of hyperparameters using a randomized search strategy with

cross-validation.

* The best model obtained from the search is stored in the variable best\_model.

Model Evaluation:

* The best model obtained is used to make predictions on the test set.
* Classification metrics such as precision, recall, F1-score, and accuracy are computed using classification\_report.
* The best hyperparameters found during the search are printed for reference.

**Departure Timeframe Prediction Initial Approach**

We started by building a neural network model for predicting the timeframe until employee departure.

This model utilized features such as age, job satisfaction, and total working years to forecast the duration an employee stays in the company.

**Departure Timeframe Prediction Revised Approach**

Due to the complexity and variability of the data, we adopted an ensemble learning technique.

This involved training multiple neural network models with dropout regularization and averaging their predictions for improved accuracy and robustness.

Data Reading and Preprocessing:

* We read the dataset.
* mapped to numeric labels (0 for 'No', 1 for 'Yes').
* select relevant features for the timeframe prediction model.

Data Splitting:

* dataset is split into training and testing sets using the (train\_test\_split) function from scikit-learn.
* The split is done with a test size of 20% and a random state for reproducibility.

Data Scaling:

* Standard scaling is applied to the training and testing features using StandardScaler.

Neural Network Model Definition:

* We define a neural network model architecture using Karas' Sequential API.
* The model consists of multiple dense layers with ReLU activation functions, followed by dropout layers to prevent overfitting.
* The output layer has a single neuron with a linear activation function, suitable for regression tasks.

Model Training:

* Multiple neural network models are trained to capture different initializations and reduce variance.
* Each model is trained for 150 epochs with a batch size of 64.
* The mean squared error loss function is used for training, along with mean absolute error, mean squared error, and accuracy as metrics.
* Training is performed on the scaled training data, with a validation split of 20% to monitor model performance during training.

Model Evaluation and Storage:

* The trained models are stored in a list for further analysis and ensemble predictions.
* Located the best evaluated model and saved for future use.

**We implemented both approaches using Python and relevant libraries such as pandas, scikit-learn, and TensorFlow.**

**Conclusion**

In conclusion, our efforts aimed to provide HR departments with predictive analytics tools to mitigate employee attrition and optimize workforce management.

By leveraging machine learning algorithms, we can empower organizations to make data-driven decisions and foster a positive work environment.

**Questions and Discussion**

Thank you for your attention. We're now open to any questions or discussions regarding our presentation and work on employee attrition prediction.