# Data Science Internship: Final Report

## Task 1

## Employee Attrition Prediction

**Task Title:** Employee Attrition Prediction

**Dataset Used:** Hr.csv

**Objective:** To predict whether an employee is likely to leave the company based on various attributes such as age, job role, department, and compensation.

**Dataset Description and Preprocessing:**

The dataset comprises 1470 entries and 35 features, covering demographic, professional, and performance-related information. Key attributes include Age, BusinessTravel, Department, DistanceFromHome, MonthlyIncome, JobSatisfaction, and more. The target variable is Attrition (Yes/No). Data cleaning involved checking for null values (none found), encoding categorical variables using one-hot encoding, and normalizing numerical columns using StandardScaler.

A black screen with white text

AI-generated content may be incorrect.

Figure 1 Dataset Inspection

A graph with a bar and a number of bars

AI-generated content may be incorrect.

Figure 2 Count for Attrition Distribution

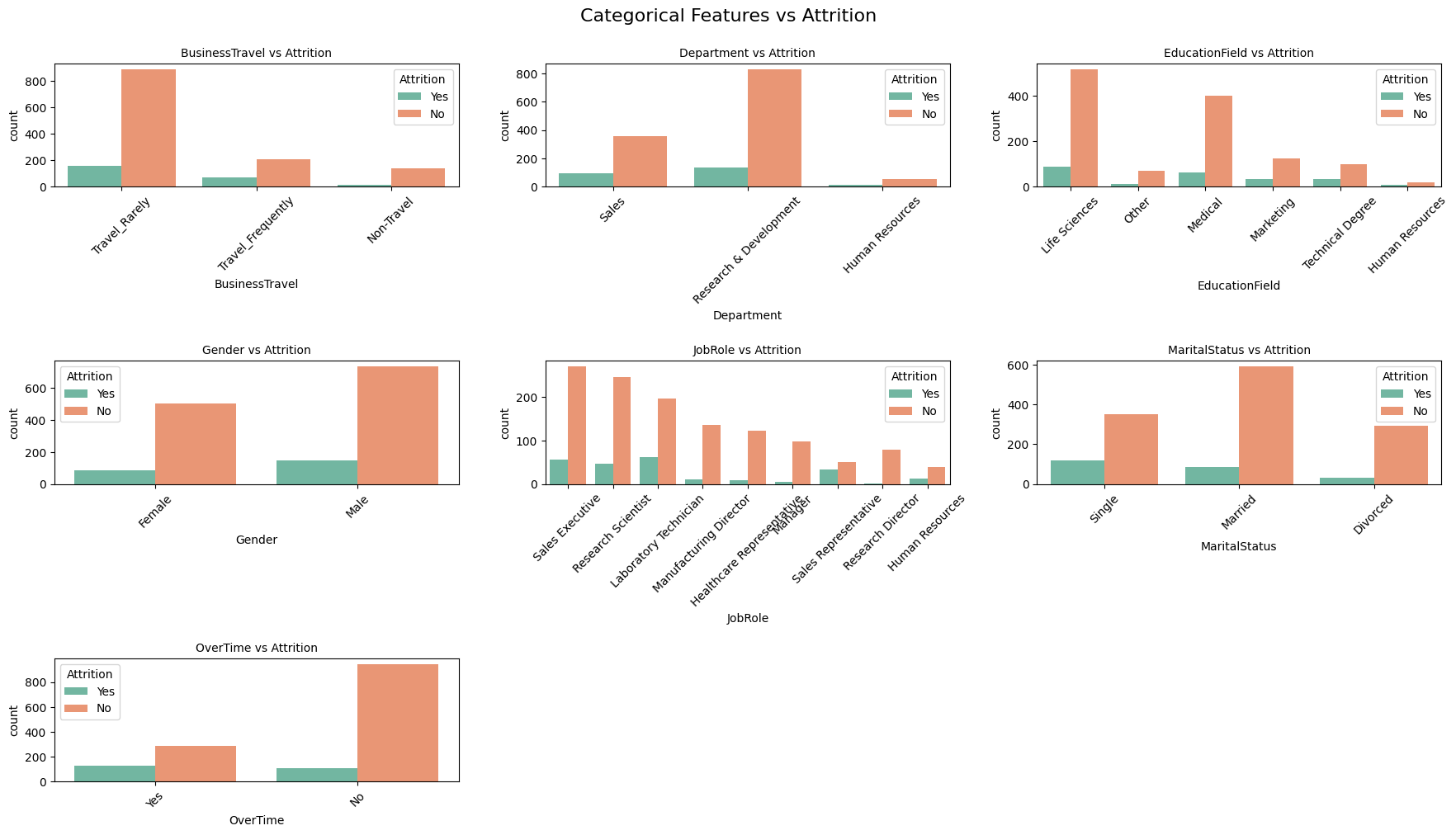


Figure 3 Categorical Features vs Attrition

**Model Implementation and Rationale:**

Three models were trained to build robust predictions:

1. Logistic Regression – selected for its simplicity and efficiency with binary classification tasks.

A screenshot of a computer

AI-generated content may be incorrect.

1. Random Forest – chosen for its ensemble nature, which handles both numerical and categorical data and reduces overfitting.

A screenshot of a computer screen

AI-generated content may be incorrect.

**Key Insights and Visualizations:**

* Attrition was more prevalent among employees in the Sales and R&D departments.
* Features such as OverTime, MonthlyIncome, and YearsAtCompany had a strong correlation with attrition.
* The Random Forest model achieved the highest performance with an AUC of 0.89.  
  Visuals included bar charts for attrition distribution, correlation heatmaps, and feature importance plots.

**Challenges and Solutions:**

**Challenge:**

While traditional performance metrics (accuracy, ROC-AUC) showed good results, model interpretability remained a challenge — especially for tree-based models like Random Forest.

**Solution:**

To tackle this, I used SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) to interpret model predictions.

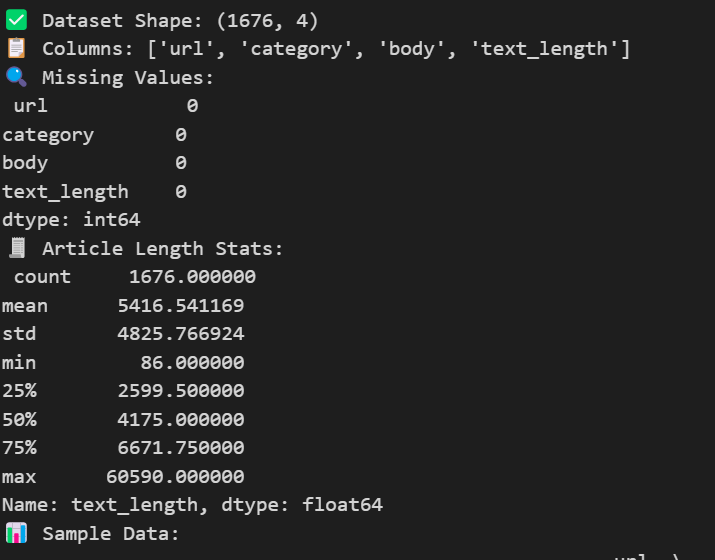
A graph of a diagram

AI-generated content may be incorrect.

# Text Summarization (NLP)

**Task Title:** News Text Summarization  
**Dataset Used**: cnn mail dataset.csv  
**Objective:** To summarize long news article texts into concise and meaningful summaries.  
  
**Dataset Description and Preprocessing:**

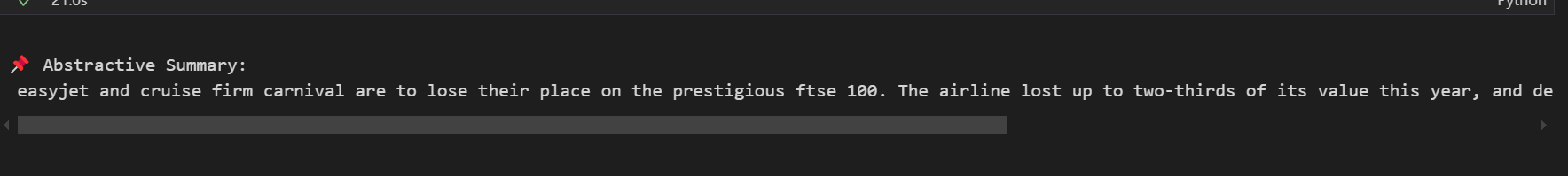
The dataset contains 1676 news articles, categorized by topics like business, politics, and health. Each row contains a URL, category, and the article body. Preprocessing included removing HTML tags, URLs, punctuation, and stop words. Tokenization and sentence segmentation were performed to structure the content.

  
  
**Model Implementation and Rationale:**1. TextRank Algorithm – an unsupervised, graph-based extractive summarization technique.  
2. T5 Transformer – a pre-trained transformer model for abstractive summarization, fine-tuned on the dataset for better accuracy.  
A graph of a distribution of text

AI-generated content may be incorrect.

**Key Insights and Visualizations:**- Word clouds revealed frequent keywords like “pandemic,” “inflation,” “policy,” indicating topical relevance.  
- Extractive summaries provided baseline performance, while abstractive methods generated more human-like narratives.

A screenshot of a computer

AI-generated content may be incorrect.  
  
**Challenges and Solutions:**  
Transformer models were resource-intensive, leading to memory issues. These were mitigated by truncating sequences and using smaller batch sizes. To reduce repetitive outputs, beam search decoding and a repetition penalty were applied.  
 

# Disease Diagnosis Prediction

**Task Title**: Diabetes Diagnosis Prediction  
**Dataset Used:** PIMA.csv  
**Objective:** To predict the likelihood of a person developing diabetes based on clinical parameters.  
  
**Dataset Description and Preprocessing:**

The dataset includes 768 samples with 9 features such as Glucose, BloodPressure, BMI, Age, and Insulin. Several fields had missing values, especially in SkinThickness and Insulin. Median imputation was used to handle these missing entries. Features were normalized to ensure consistency across different scales.

A screenshot of a computer screen

AI-generated content may be incorrect.

A group of blue and black graphs

AI-generated content may be incorrect.

Figure 4 Boxplot

A screenshot of a computer screen

AI-generated content may be incorrect.

Figure 5 Correlation Map

**Model Implementation and Rationale:**

1. Gradient Boosting – used for its interpretability and as a baseline classifier.
2. K-Nearest Neighbors – leveraged for its non-parametric nature and simplicity.
3. SVM – chosen for their ensemble learning capabilities, improving accuracy and stability.  
   **Key Insights and Visualizations:**

- High glucose levels and BMI were the most significant indicators of diabetes.  
- The XGBoost model provided the best classification accuracy (around 82%) and ROC-AUC of 0.88.  
- Visuals included histograms, boxplots for feature distributions, and confusion matrices for model performance.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

**Challenges and Solutions:**

A high percentage of missing values in Insulin and SkinThickness was a major hurdle. These were addressed using statistical imputation. Outliers were handled by minorizing the feature distributions.

A screenshot of a computer

AI-generated content may be incorrect.

# Loan Default Prediction

**Task Title:** Loan Default Risk Prediction  
**Dataset Used:** loan.csv  
**Objective:** To predict whether a loan applicant will default based on their financial and employment history.  
  
**Dataset Description and Preprocessing:**

The dataset contains 20,000 entries and 15 features, including income, credit history, purpose, employment length, and loan grade. Missing values were present in fields like home\_ownership and last\_major\_derog\_none. Categorical features were encoded using LabelEncoder or one-hot encoding, while numerical features were scaled.

A screen shot of a computer

AI-generated content may be incorrect.

A screenshot of a computer screen

AI-generated content may be incorrect.

A group of graphs with numbers

AI-generated content may be incorrect.  
  
**Model Implementation and Rationale:**

* LightBGM
* SVM

A computer screen with white text

AI-generated content may be incorrect.  
**Key Insights and Visualizations:**

* Default risk increased with higher debt-to-income ratios and shorter employment history.
* Customers with purposes like “debt\_consolidation” had a significantly higher likelihood of default.
* AUC score was around 0.85 with XGBoost.
* Graphs included ROC curves, default distributions by loan purpose, and feature importance plots.  
  A screenshot of a computer

  AI-generated content may be incorrect.  
  **Challenges and Solutions:**

Handling imbalance in the target variable was critical and was addressed by applying balanced class weights. Missing values in categorical columns were imputed with the most frequent category. Outliers in numerical features like annual income and revol\_util were capped.

A graph with numbers and a bar

AI-generated content may be incorrect.