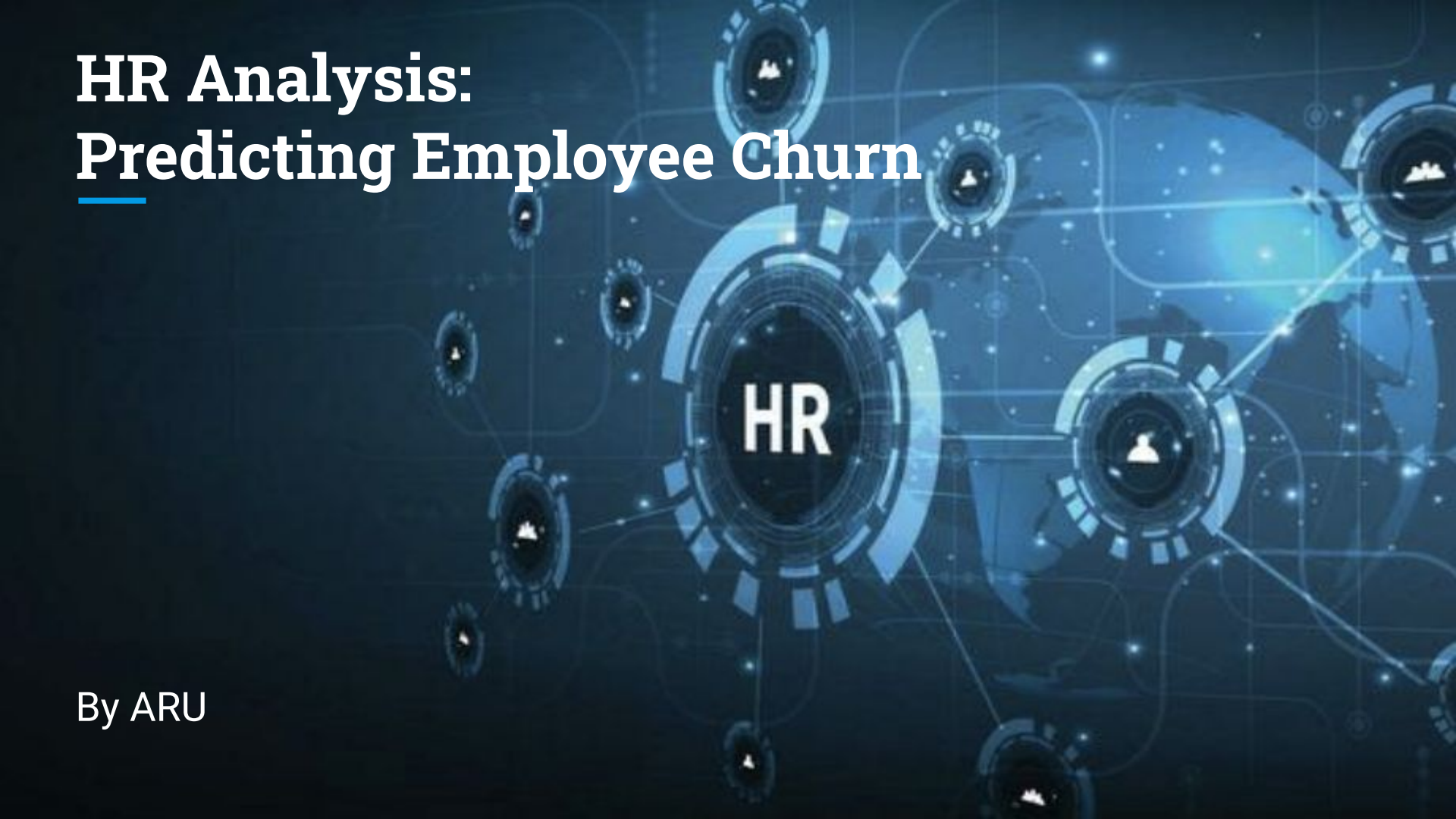
HR Analysis: Predicting Employee Churn



By ARU

HR Analysis:
Predicting Employee Churn

- **What is it about:** The workforce demographics, the resignation ratios, and developing a resignation prediction model
- **Why did I choose it:** Because of my best friend
- **Project Goal:** To visualize the insights, build the predictive model and apply as many skills as possible learned throughout the bootcamp

Methodology

Data collection and Preparation

Data Collection: Kaggle

Data Storage: SQL, retrieved into pandas DataFrame using SQL alchemy

Data Cleaning & Formatting: Handled missing values, outliers and ensured data consistency

Feature engineering: Applied **label encoding** and **ordinal encoding** for categorical features

Scaling: Applied **StandardScaler** to numerical features for models like **logistic regression**

Methodology

EDA

Methodology

Model selection - Classification

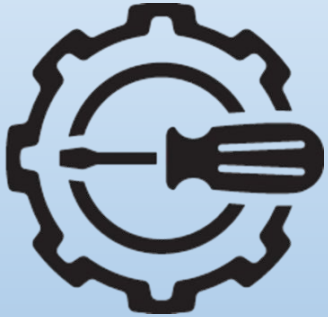
Features selected: Gender, Age, Income, Work environment, Job title, Performance, Satisfaction score, Overtime hours, Training hours and Promotions

- Logistic Regression
- Decision Tree
- Random Forest
- Ada Boosting
- XGBoost
- Catboost
- Lightboost

- XGBoost

- Effective with imbalance data
- High predictive power
- Fast with large datasets
- Supports mixed data
- Supports customized metrics

Methodology



Hyperparameter Tuning

- Random Search
- Grid search
- Cross Validation

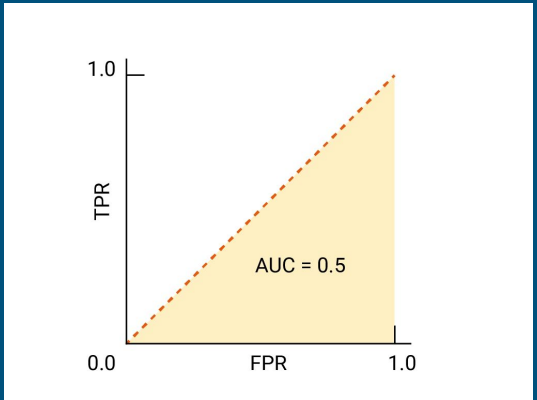
```
param_grid = {
    'learning_rate': [0.01, 0.05, 0.1],
    'n_estimators': [50, 100, 200],
    'max_depth': [3, 5, 7],
    'scale_pos_weight': [1, 5, 9]}
```

* SMOTE resampling technique was used in the pipeline to tackle the highly imbalanced data

Methodology

Model Evaluation

- Accuracy: 0.9
- Recall: 0.76
- ROC-AUC: 0.52

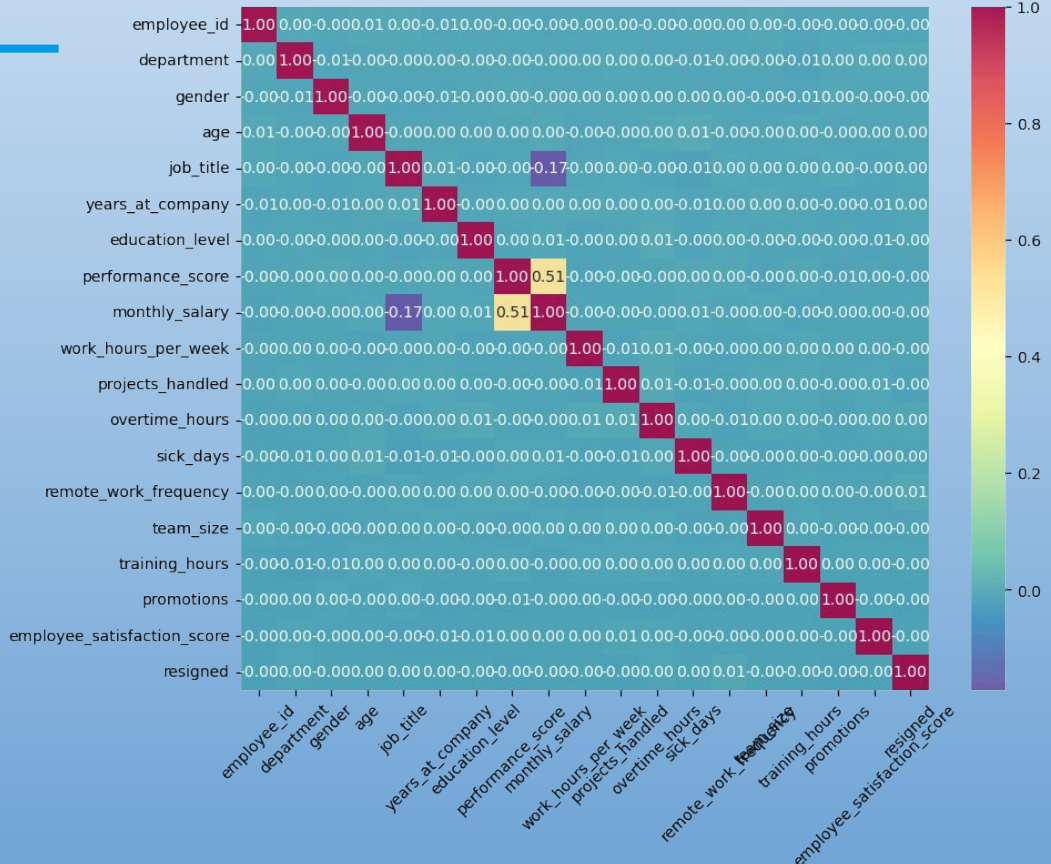


ROC Curve: The ROC curve is a plot that shows the True Positive Rate (Recall) on the Y-axis against the False Positive Rate on the X-axis for different threshold values. Essentially, it shows how well the model can separate the positive class from the negative class across all possible decision boundaries.

Why ROC-AUC?

ROC-AUC focuses on the model's ability to **separate the two classes** (resigned vs. not resigned) rather than simply counting correct predictions (accuracy).

Model Evaluation: Possible Reasons



- There is no significant correlation, except for salary being correlated with job title and performance score
- The synthetic data might be poorly constructed and may not accurately reflect reality

Conclusion and Future Work

- Incorporate Additional Features
- Integrating more employee data, such as workload indicators, and engagement survey results, which might provide deeper insights into factors influencing resignation.
- Incorporating external data like industry trends, economic conditions, or market salary data, which may impact employee decisions
- Experiment with Additional Algorithms: Experimenting with advanced models or techniques like deep learning if more data becomes available or ensembles of multiple models to improve predictive accuracy
- The synthetic data might not capture the real world data features
- Find other real word data :)

THANK YOU!

By ARU