[redacted] HR

x

CU Leeds

Spring 2024, Group 6
Honor Brogden – hobr2220@colorado.edu, Megha Gupta – megu2186@colorado.edu,
Nathan Kareithi – naka6894@colorado.edu, Jackson Miers – jami1731@colorado.edu,
Maddie Wallace – mawa9164@colorado.edu, Fizza Zaidi – fiza2251@colorado.edu

Table of contents

01

Business Understanding & Problem Statement

Problem Statement...

Challenges

Difficulties with efficient data integration and analysis across multiple systems.

Current Limitations

Existing tools like Excel are time-consuming and lack repeatability.

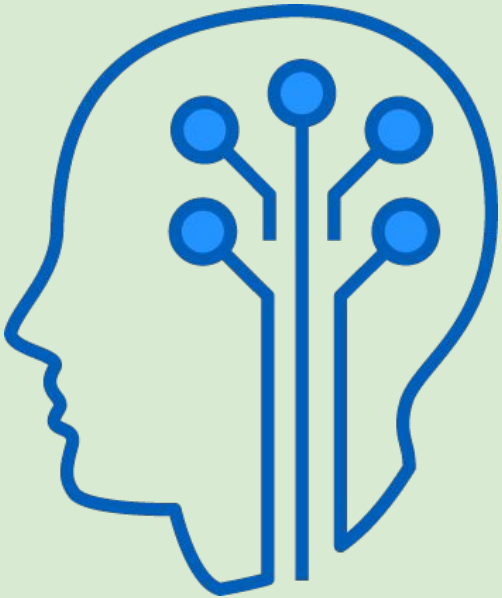
Predictive Goals

Identify drivers of voluntary churn, assess financial impact, and forecast future human capital needs.

What you need from us...

[redacted] HR seeks to:

- Enhance analysis capabilities
- Incorporate predictive analytics and ML
- Improve timeliness



02

Overarching Objectives

What we can do for you...

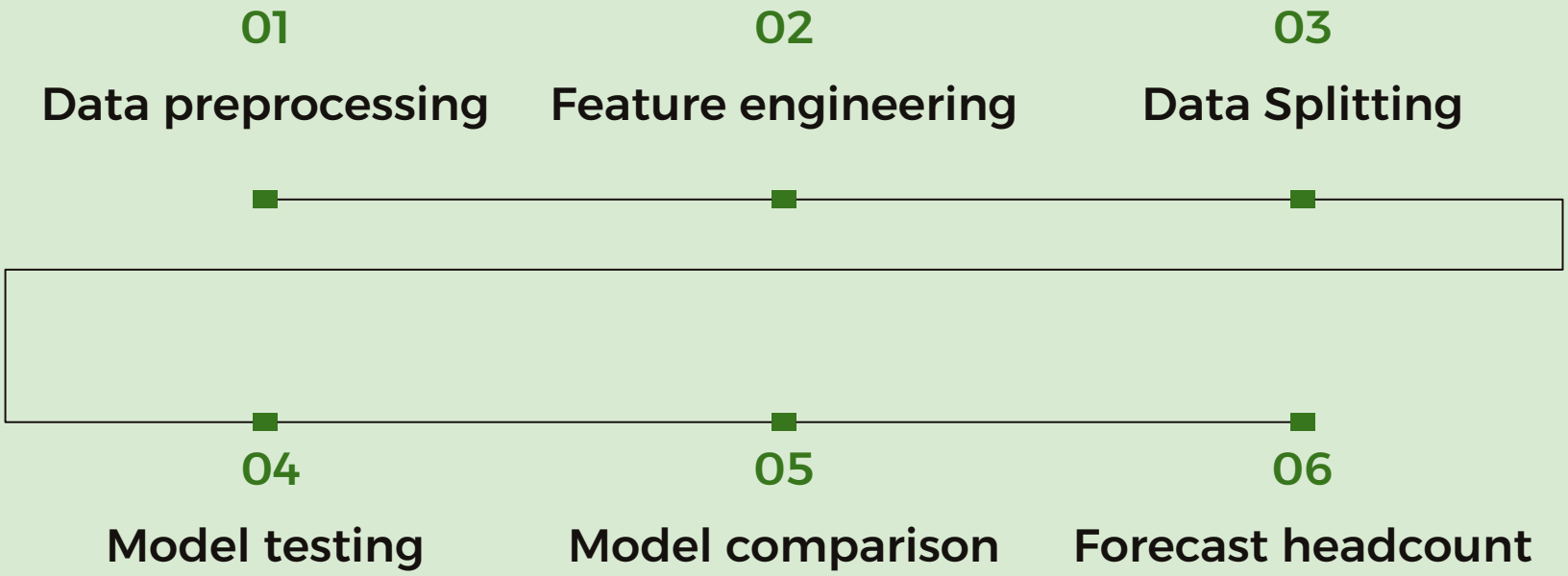
Integrate, clean, analyze data

Identify key drivers of voluntary churn

Improve efficiency and repeatability

Forecast human capital

Overview of Process



03

Data Cleaning & Set Up

Preprocessing

01. Column Names

- All lower case
- ‘anon id’
- ‘termination reason’

02. Drop NA’s

- generation
- pay level
- currency conversion rate

03. Drop Duplicates

Feature Engineering

01. Work Location

- ‘work city’
- ‘british columbia’
- ‘Onsite’ structure

02. Tenure

- Rounded to 2 decimal places
- Binned into 5 different bins
- '<1', '1–5', '5–10', '10–20', '20+'

03. Cost to Replace Employee

- base pay mid point annualized' x 'cost to replace employee multiplier'

04. USD mid point

- 'base pay mid point annualized' to USD using 'currency conversion rate'

05. Voluntary Churn

- Marked whether employee had churned voluntarily or not

06. Compa Ratio

- Removed outliers
- Binned into 5 different bins
- 0, 0.25, 0.5, 0.75, 1

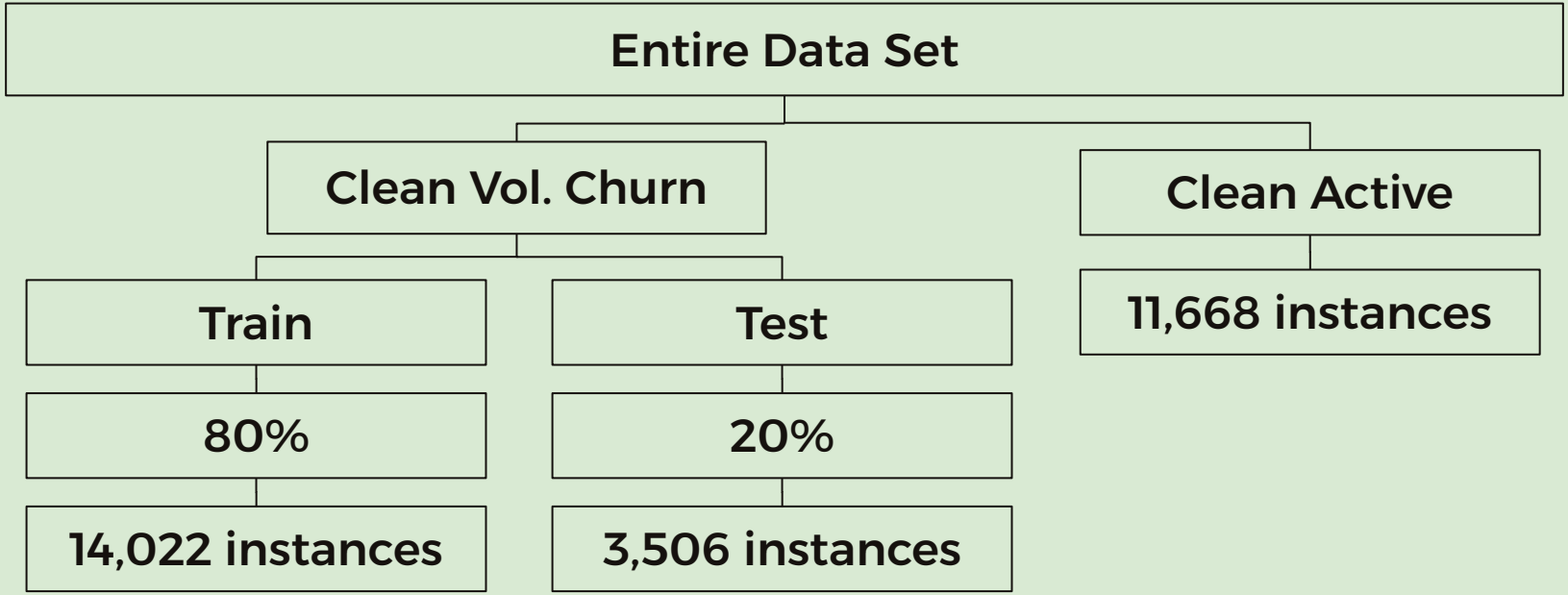
07. One-Hot Encoding

- converts categorical variables into binary vectors
- each category represented by a separate binary column
- Vital for algorithms to understand data

08. RFE

- determined the top 100 most relevant features for one of our high-performing models, and used those 100 features as our X variables for the remaining models

Data Splitting



04

Models Used

Model Types Used

Gradient Boosting Classifier

Often yields high performance and robustness to noisy data

Support Vector Machine

Effective in handling high-dimensional data and is robust to overfitting

Random Forest Classifier

Effective in handling high-dimensional data and capturing complex relationships within the dataset

eXtreme Gradient Boosting Classifier

Robustness and superior handling of varied data types

Logistic Regression

Robust, interpretable, and suitable for scenarios with linearly separable data

05

Model Comparison

Metrics Explained

Accuracy

The proportion of correctly classified instances out of the total instances

Recall

The proportion of true positive predictions among all actual positive instances

F1

Harmonic mean of precision and recall

AUC

The area under the receiver operating characteristic (ROC) curve, which plots the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings

	Model	Accuracy	F1-Score	Recall	AUC	Choice
0	GBC	78.18%	62.95%	55.46%	72.52%	XX
1	SVC	78.78%	62.27%	52.38%	72.21%	XX
2	RFC	78.69%	65.56%	78.69%	82.89%	✓
3	XGB	79.78%	66.91%	61.17%	84.72%	✓
4	LR1	77.21%	61.11%	53.58%	71.33%	XX
5	LR2	77.23%	61.19%	53.66%	71.37%	XX

Accuracy: how well the model is performing across all classes
F1: provides a balanced measure that considers both precision and recall. Gives equal weight to both false positives and false negatives
Recall: captures all positive instances, indicating how many of the actual positive instances are correctly identified
AUC: represents the model's ability to distinguish between positive and negative classes. A higher AUC value indicates better discrimination performance.
Choice: whether we moved forward with the model

In-Depth Comparison, Top 3 Models

A) Support Vector Machine

Accuracy: ~ 79% of the predictions are correct
F1-score: overall accuracy, considering both precision and recall, is ~ 62%
Recall: Correctly identifies ~ 52% of all actual positive instances
AUC: ability to distinguish between positive and negative classes is ~ 72%

B) XGBoost Classifier

Accuracy: ~ 80% of the predictions are correct.
F1-score: overall accuracy, considering both precision and recall, is ~ 67%
Recall: correctly identifies ~ 61% of all actual positive instances
AUC: ability to distinguish between positive and negative classes is ~ 85%

C) Random Forest Classifier

Accuracy: ~ 79% of the predictions are correct.
F1-score: overall accuracy, considering both precision and recall, is ~ 66%
Recall: correctly identifies ~ 79% of all actual positive instances
AUC: ability to distinguish between positive and negative classes is ~ 83%.

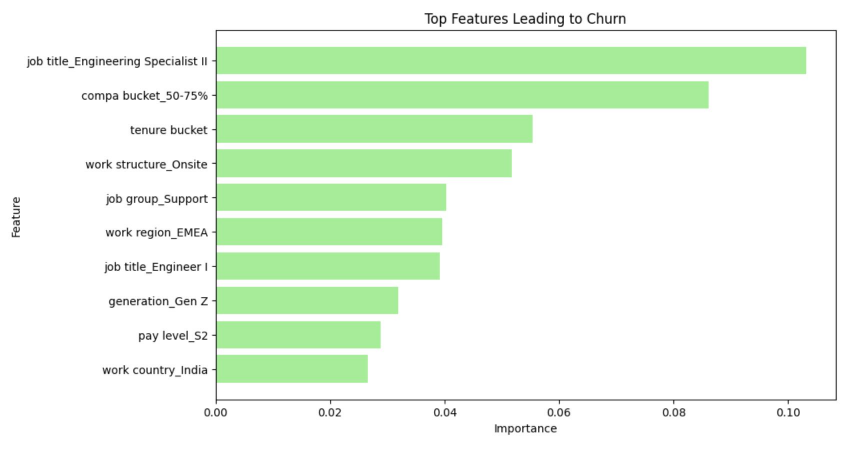
06

Final Suggestion

Final Recommendation

Benefits of Adopting XGBoost Classifier

- **Proactive Churn Management:** Helps identify and address key drivers of churn
- **Reduced Turnover Rates:** Potential to decrease turnover rates and improve employee satisfaction.
- **Optimized Recruitment Strategies:** Insights from model assist in refining recruitment strategies, aligning with forecasted human capital needs.



Final Recommendation

Implement XGBoost Classifier

- **Model Choice:** XGBoost Classifier (80% accuracy)
- **Performance Metrics:** Highest accuracy(80%), F1-score(67%), and AUC(85%)
- **Prediction and Classification:** Demonstrates superior capability in predicting and classifying potential voluntary churn

	Predicted Active	Predicted Voluntary Termination
Actual Active	2080	254
Actual Voluntary Termination	455	717

Forecasted Headcount

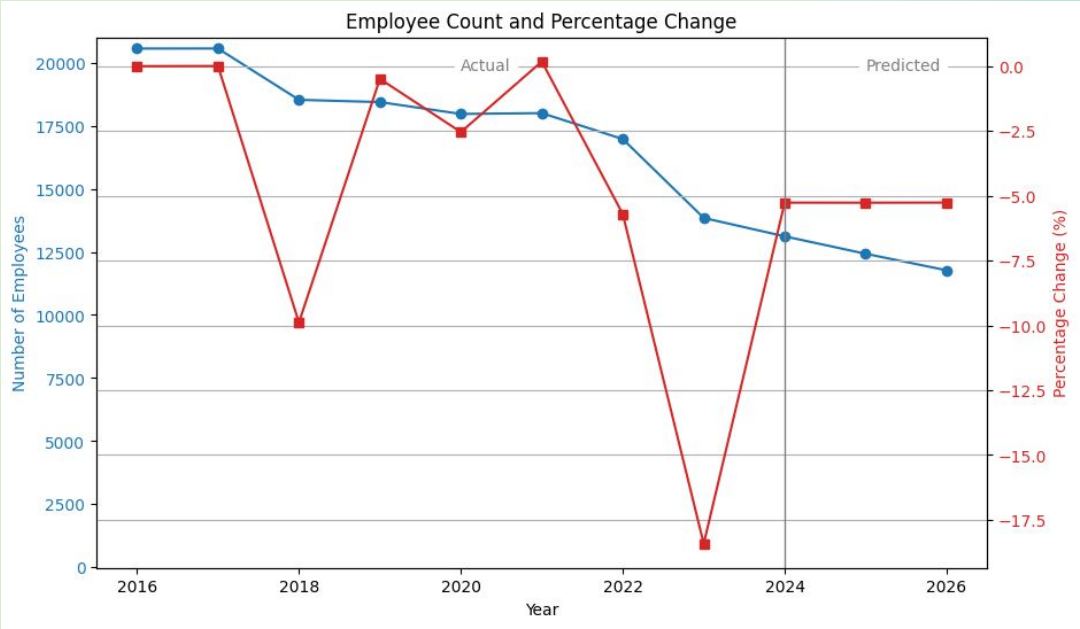
Using our XGB model on the data with current active employees, we were able to predict how many of the active employees are likely to churn, as well as the churn probability for every active employee.

	Likely to Churn
False	11,431
True	237

Anon ID	Churn Probability	Likely to Churn
1113	02.32%	False
1115	13.43%	False
1116	21.88%	False
1120	01.47%	False

Forecasted Headcount Cont.

- 5.27% employee loss / year
- hire a number of employees ~ equal to 6% of their current headcount



Thank you!

Honor Brogden – hobr2220@colorado.edu,
Megha Gupta – megu2186@colorado.edu,
Nathan Kareithi – naka6894@colorado.edu,
Jackson Miers – jami1731@colorado.edu,
Maddie Wallace – mawa9164@colorado.edu,
Fizza Zaidi – fiza2251@colorado.edu