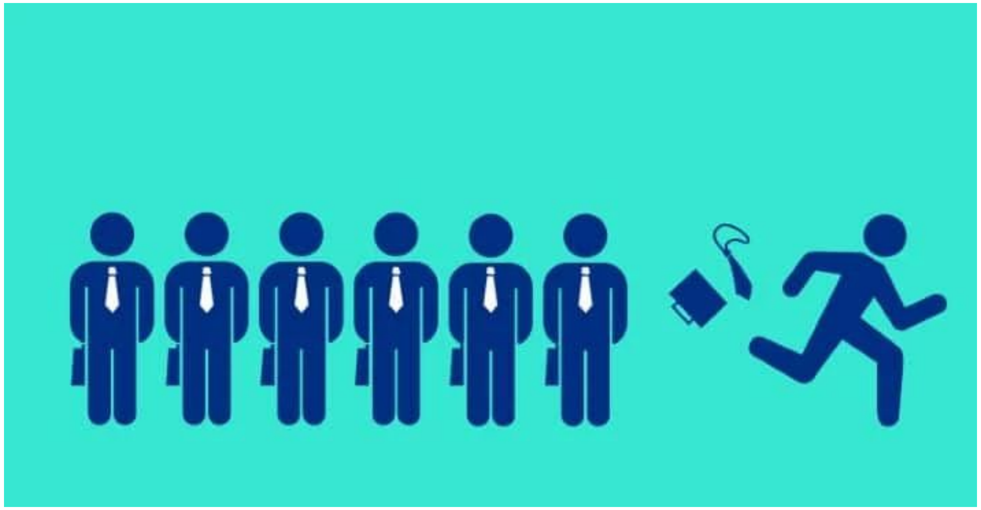


Predicting Employee Attrition Using Logistic Regression, Decision Tree, and KNN

University of California, Berkeley | School of Information

DATASCI 207 Applied Machine Learning - Spring 2023

Final Project | Final Presentation

Team: Ivy Chan, John Gibbons, Mark Herrera, Maria Manna

April 18, 2023

Berkeley
UNIVERSITY OF CALIFORNIA

AGENDA

1. Motivation & Research Question
2. Data
3. Approach
4. Experiments
5. Conclusions

Berkeley
UNIVERSITY OF CALIFORNIA

Motivation & Research Question

Research Question

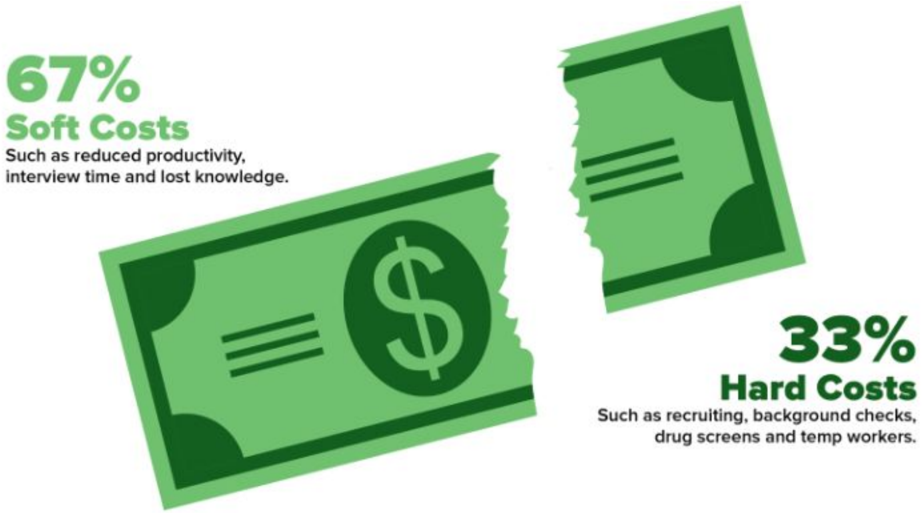
- Can we use employee characteristics (age, department, compensation, etc.) to accurately predict attrition?
- What factors are most predictive of an employee's attrition?

Understanding these questions allows businesses to better identify at-risk employees. It can also assist in the hiring process if there are known factors that influence employee churn. In the past two years, employers have been significantly impacted by events such as the “Great Resignation,” losing employees to other employment opportunities. The knowledge gained by our research has the potential to lessen the impact of future periods of high employee turnover for businesses that value a low employee turnover rate.

Berkeley
UNIVERSITY OF CALIFORNIA

Motivation

- US businesses lost ~$1 Trillion to voluntary turnover in 2022 (up from ~$600B in 2018)
- The BLS reported 57% overall turnover rate for 2022
  - The overall annual turnover rate includes voluntary (quits), involuntary, and other (retirement, death) separations, with voluntary accounting for 70-75% of total separations
  - 50.6 million quits - highest in JOLTS history
- 87% of all workers are open to new job opportunities (LinkedIn)
- 52% of voluntarily exiting employees say their manager or organization could have done something to prevent them from leaving their job (Gallup)



Motivation

Position Type	Average Replacement Cost
Entry-level/non-skilled	30-50% of employee's annual salary
Service/production	40-70% of employee's annual salary
Clerical/administrative	50-80% of employee's annual salary
Skilled hourly	75-100% of employee's annual salary
Professional	75-125% of employee's annual salary
Technical	100-150% of employee's annual salary
Supervisor	100-150% of employee's annual salary

https://www.gnapartners.com/resources/articles/how-much-does-employee-turnover-really-cost-your-business

Berkeley
UNIVERSITY OF CALIFORNIA

Value Proposition

- What's interesting is that while there are many consulting firms specializing in compensation techniques, recruiting, HR, staffing services, "talent solutions" etc., finding a business that primarily focuses on employee retention was a lot easier said than done. The primary application of most were either recruiting or human resources, with the potential to add other packages that could help identify areas of improvement for retention

__MM to work__

If we can do this well…

Berkeley
UNIVERSITY OF CALIFORNIA

Overall Plan & Summary of Results

1. **Utilize Logistic Regression as the Baseline Model**
   a. Utilize upsampling and downsampling to account for class imbalance

2. **Experiment Model 1 - Decision Tree**

3. **Experiment Model 2 - K-Nearest Neighbors (KNN)**

Findings:

- **Baseline Model: Logistic Regression**
  - Maintained the highest recall score of the three models
- **Model #1: Decision Tree**
  -
- **Model #2: KNN**
  - Achieved highest accuracy score of the three models

Berkeley
UNIVERSITY OF CALIFORNIA

Data

Data

We will use an employee attrition dataset from Kaggle containing the employee data from IBM HR Employee Attrition and Performance. The data from this dataset is structured.

The dataset size: 1,470 observations and 35 features

The target variable: Attrition (binary - 1 for attrit, 0 otherwise)

Features chosen for analysis:

- Age
- EnvironmentSatisfaction
- JobSatisfaction
- StockOptionLevel
- WorkLifeBalance
- departmentResearch&Development
- MonthlyIncome

Summary Statistics

	Attrition	Age	EnvironmentSatisfaction	JobSatisfaction	StockOptionLevel	WorkLifeBalance	departmentResearch&Development	MonthlyIncome
count	940.000	940.000	940.000	940.000	940.000	940.000	940.000	940.000
mean	0.162	37.076	2.706	2.763	0.816	2.753	0.657	6563.973
std	0.368	9.154	1.118	1.092	0.844	0.723	0.475	4700.521
min	0.000	18.000	1.000	1.000	0.000	1.000	0.000	1051.000
25%	0.000	30.000	2.000	2.000	0.000	2.000	0.000	2909.000
50%	0.000	36.000	3.000	3.000	1.000	3.000	1.000	4964.000
75%	0.000	43.000	4.000	4.000	1.000	3.000	1.000	8634.500
max	1.000	60.000	4.000	4.000	3.000	4.000	1.000	19999.000

*df = training data subsetted for key variables of interest (data not yet standardized)

Summary Statistics

	Attrition	Age	EnvironmentSatisfaction	JobSatisfaction	StockOptionLevel	WorkLifeBalance	departmentResearch&Development	MonthlyIncome
count	940.000	940.000	940.000	940.000	940.000	940.000	940.000	940.000
mean	0.162	37.076	2.706	2.763	0.816	2.753	0.657	6563.973
std	0.368	9.154	1.118	1.092	0.844	0.723	0.475	4700.521
min	0.000	18.000	1.000	1.000	0.000	1.000	0.000	1051.000
25%	0.000	30.000	2.000	2.000	0.000	2.000	0.000	2909.000
50%	0.000	36.000	3.000	3.000	1.000	3.000	1.000	4964.000
75%	0.000	43.000	4.000	4.000	1.000	3.000	1.000	8634.500
max	1.000	60.000	4.000	4.000	3.000	4.000	1.000	19999.000

*df = training data subsetted for key variables of interest (data not yet standardized)

Berkeley
UNIVERSITY OF CALIFORNIA

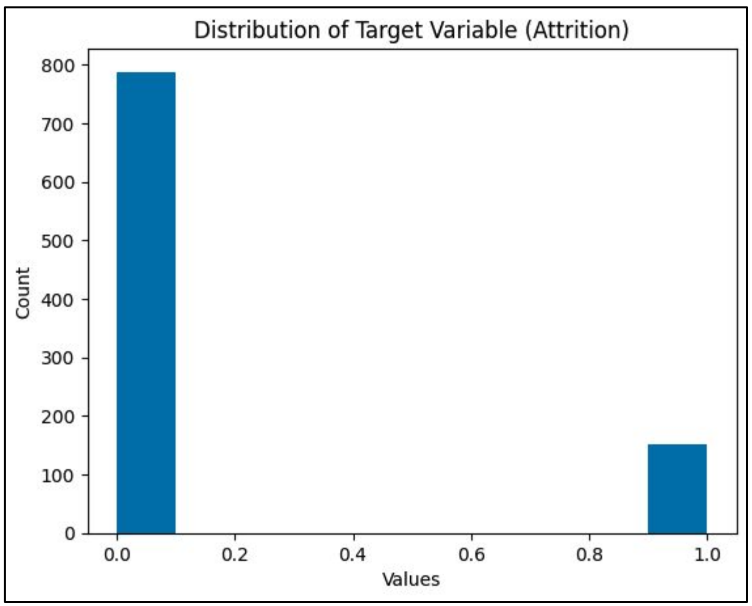
Scaling Method & Class Imbalance

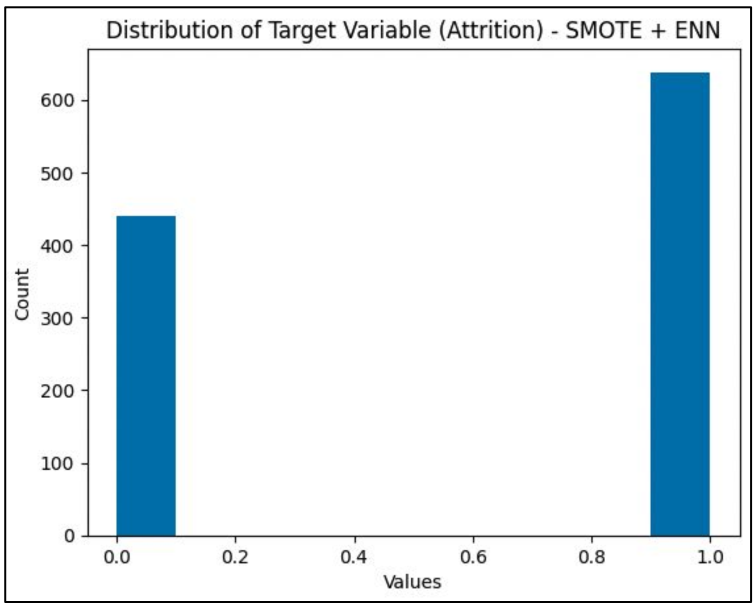
Scaling Method	We used standardization to scale the training and validation datasets to make sure all data points are measured and represented in a consistent and uniform way.
Class Imbalance	We implemented a hybrid upsampling/downsampling technique referred to as **SMOTE + ENN** address the significant class imbalance.

- **SMOTE** (*Synthetic Minority Over-sampling Technique*) generates synthetic samples by interpolating new points between existing minority class examples.
- This sometimes generates noisy samples that do not accurately represent the minority class. SMOTE can be combined with **ENN** (*Edited Nearest Neighbours*), which removes noisy samples by analyzing the k-nearest neighbors of each sample.
- The **SMOTE + ENN** technique first applies SMOTE to oversample the minority class, then uses ENN to remove noise. If the majority of the k-nearest neighbors belong to a different class than the sample being examined, that sample is considered noisy and removed.

Scaling Method & Class Imbalance

SMOTE + ENN	Training Dataset # of Row	Validation Dataset # of Row	% of Attrition (training data)
Before upsampling/downsampling	940	236	16%
After upsampling/downsampling	1078	278	59%





Approach

Overall Plan: Prediction Algorithms

1. **Baseline Model - Logistic Regression**
   a. Use: binary classification problems (will an employee turnover or not?)
   b. Will utilize Stochastic Gradient Descent (SGD)

2. **Experiment Models 1 - Decision Tree + Random Forest**
   a. Uses: feature selection & prediction
   b. Measure with information gain

3. **Experiment Model 2 - K-Nearest Neighbors (KNN)**
   a. Use: identify employees with higher risk of attrition by comparing to profiles of known former employees
   b. Measure with manhattan distance

Berkeley
UNIVERSITY OF CALIFORNIA

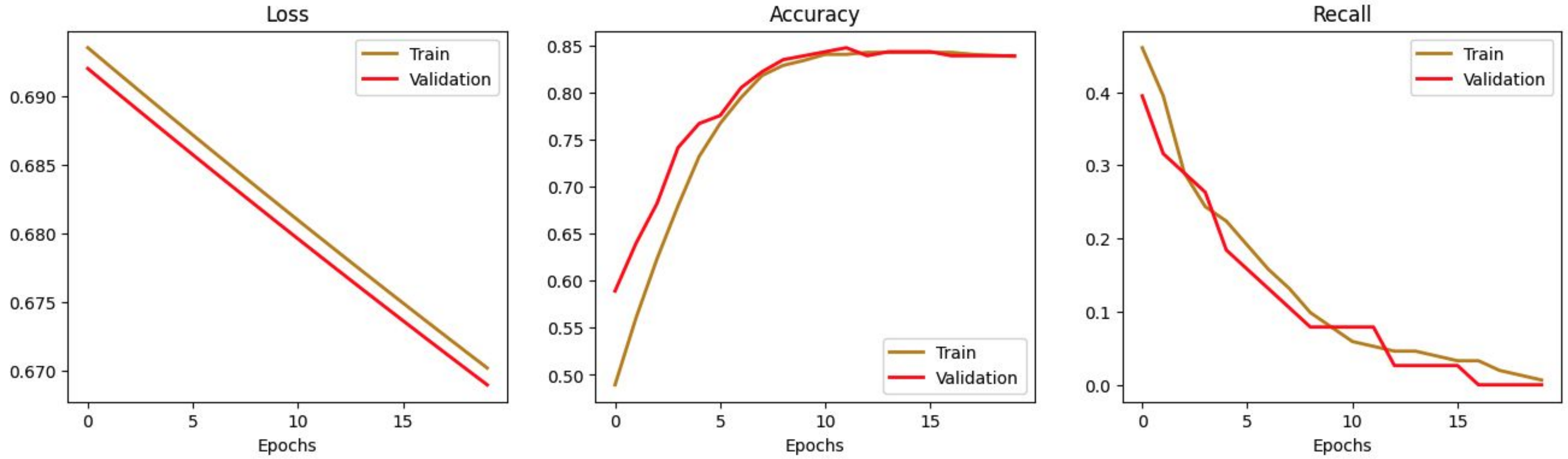
Baseline Model - Logistic Regression

Model basics:

- Activation: Sigmoid
- Loss: Binary Crossentropy
- Optimizer: SGD (learning_rate=0.01)
- Metrics: Accuracy, Recall
- Epochs: 20

Initial results:

- Decent accuracy (0.8389 after 20 epochs, which is comparable to guessing 0 for all examples)
- However, this was at the expense of recall
- Insight: without upsampling, the model achieves decent accuracy by ignoring the minority class



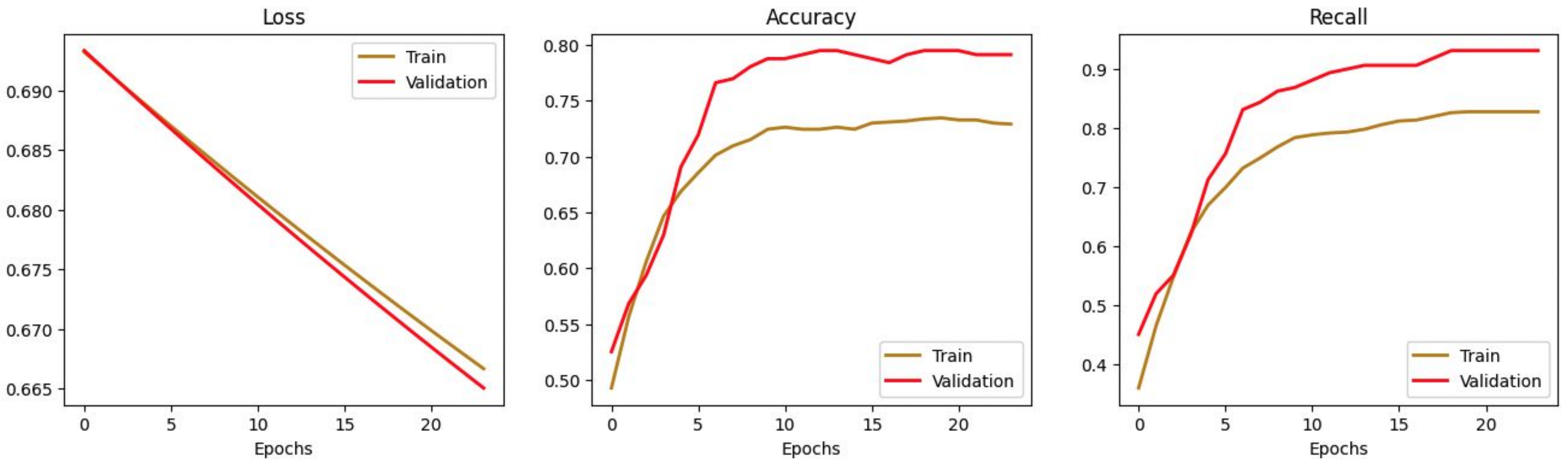
Baseline Model - Logistic Regression with SMOTE + ENN

Model basics:

- Activation: Sigmoid
- Loss: Binary Crossentropy
- Optimizer: SGD (learning_rate=0.01)
- Metrics: Accuracy, Recall
- Epochs: 50 + callbacks

Initial results:

- Slightly lower accuracy than non-upsampled (0.7913) but dramatic improvement in recall (0.9312) after 23 epochs
- Insight: SMOTE + ENN upsampling to correct for class imbalance appear successful - will use this method for all models



Baseline Model - Evaluation on Test Data

Data	Accuracy	Recall	# epochs
Original data	0.8389	0.0000	20
SMOTE + ENN	0.7913	0.9312	23
Test data	0.4863	0.8510	n/a

Key takeaways:

- Model maintained a decent recall score, but saw a dramatic decrease in accuracy, potentially indicating the opposite problem from the non-upsampled data (over-labeling of 1 instead of over-labeling of 0)
- Clear room for improvement in the subsequent models

Berkeley
UNIVERSITY OF CALIFORNIA

Experiments

Experiment Model 1 - Decision Tree

- <u>Decision Tree Model</u> - splits data on features that maximize information gain, recursively selecting optimal features that meet this criteria until each class is a separate leaf node or a maximum depth is reached.
  - Hyper parameters tuned: tree depth

Berkeley
UNIVERSITY OF CALIFORNIA

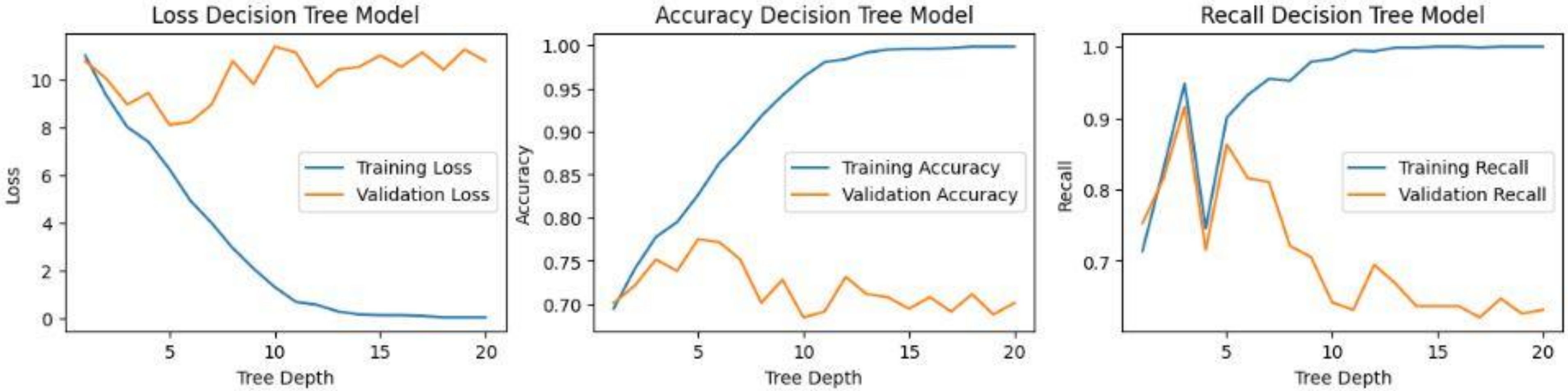
Model 1 - Decision Tree with SMOTE + ENN

Model basics:

- Tree Depth: 5
- Criterion: entropy

Initial results:

- Accuracy was relatively high at 75% but recall was low with 40%.
- Insight:
  - The model seems to be primarily predicting the majority class (no attrition)
  - Features selected to split on in order or importance (stock option level, Overtime, Department - Research and Development, Job Role - Lab Tech, Education - Technical degree)



Experiment Model 2 - KNN Hyperparameter Tuning

• Lazy learner	• No training is involved	• Memorizes data

Hyperparameter Tuning on Validation Dataset:

```
KNN hyperparameters tuning using GridSearchCV:
----------------------------------------------
Best hyperparameters:  {'metric': 'manhattan', 'n_neighbors': 3, 'weights': 'distance'}
Accuracy: 0.978
```

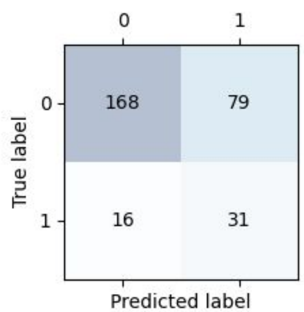
Fit Model on Training Dataset & Predict Label on Validation Dataset:

```
KNN log loss for validation set: 8.427
KNN accuracy score for validation set: 0.766
```

Berkeley
UNIVERSITY OF CALIFORNIA

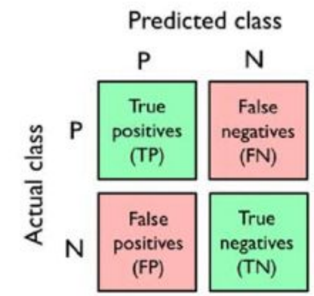
Experiment Model 2 - KNN Confusion Matrix & Classification Report

Prediction Result on Test Dataset



Classification Report for KNN Model on Test Set:

	precision	recall	f1-score	support
0	0.91	0.68	0.78	247
1	0.28	0.66	0.39	47
accuracy			$\boxed{0.68}$	294
macro avg	0.60	0.67	0.59	294
weighted avg	0.81	$\boxed{0.68}$	0.72	294



$$Accuracy = \frac{TN + TP}{TN + FP + TP + FN}$$

$$Recall = \frac{TP}{TP + FN}$$

Key takeaways:

- Difficult to interpret how the model is making predictions.
- Favor the majority class as it simply picks the most frequent class among the K nearest neighbors in imbalance class distribution.
- Need to select the value of K carefully: a low value of K may lead to overfitting, and a high value of K may lead to underfitting.

Berkeley
UNIVERSITY OF CALIFORNIA

Conclusions

Result Summary

Model	Accuracy	Recall	Analysis
Logistic Regression	48%	**85%**	• Highest recall among all models • High recall, low accuracy • Without upsampling, overpredicted majority class; with it, overpredicted the minority class
Decision Tree	**75%**	40%	• Highest accuracy among all models • Lowest Recall among all models • Over Predicting majority class (no attrition). Possible source is difference in distributions of data sets.
KNN	68%	68%	• Same accuracy & recall • Made roughly equal numbers of true positive and true negative predictions, while also made some false positive and false negative predictions. • Should also evaluate with other metrics

Berkeley
UNIVERSITY OF CALIFORNIA

Thank You

References

GitHub Repo & Dataset Links

Name	Link
GitHub Repo	https://github.com/ivykamanchan/207_Final_Project
Dataset	https://www.kaggle.com/code/hamzaben/employee-churn-model-w-strategic-retention-plan/data

Note: The GitHub Repo is made public to ensure the instructor can access.

Berkeley
UNIVERSITY OF CALIFORNIA

Contributions

Tasks	Team Members
Topic Research	Ivy Chan, John Gibbons, Mark Herrera, Maria Manna
Dataset Review	Ivy Chan, John Gibbons, Mark Herrera, Maria Manna
Data Processing	Ivy Chan, John Gibbons, Mark Herrera, Maria Manna
Algorithm Selection & Implementation	Ivy Chan, John Gibbons, Mark Herrera, Maria Manna
Code Review	Ivy Chan, John Gibbons, Mark Herrera, Maria Manna
Slides & Presentation	Ivy Chan, John Gibbons, Mark Herrera, Maria Manna
Meetings & Discussions	Ivy Chan, John Gibbons, Mark Herrera, Maria Manna

References

Content

- https://scikit-learn.org/stable/supervised_learning.html
- https://medium.com/@mikeusru/common-metrics-for-evaluating-natural-language-processing-nlp-models-e84190063b5f
- https://xgboost.readthedocs.io/en/stable/get_started.html
- https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62
- https://www.v7labs.com/blog/f1-score-guide#:~:text=F1%20score%20is%20a%20machine%20learning%20evaluation%20metric%20that%20measures,prediction%20across%20the%20entire%20dataset.
- https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
- https://towardsdatascience.com/gridsearchcv-for-beginners-db48a90114ee
- https://www.terrastaffinggroup.com/
- https://www.gnapartners.com/resources/articles/how-much-does-employee-turnover-really-cost-your-business
- https://www.peoplekeep.com/blog/employee-retention-the-real-cost-of-losing-an-employee
- https://www.gallup.com/workplace/247391/fixable-problem-costs-businesses-trillion.aspx
- https://www.apollotechnical.com/employee-retention-statistics/#:~:text=As%20a%20general%20rule%2C%20employee,rate%20of%2010%25%20or%20less
- https://www.bls.gov/jlt/home.htm
- https://www.bls.gov/news.release/pdf/jolts.pdf
- https://www.business.com/articles/employee-turnover-rate/
- https://www.shrm.org/hr-today/trends-and-forecasting/research-and-surveys/Documents/2017-Human-Capital-Benchmarking.pdf
- https://stats.stackexchange.com/questions/99694/what-does-it-imply-if-accuracy-and-recall-are-the-same
- https://vitalflux.com/micro-average-macro-average-scoring-metrics-multi-class-classification-python/#:~:text=Use%20macro%2Daveraging%20score%20when,related%20to%20different%20class%20labels).
- https://medium.com/analytics-vidhya/confusion-matrix-accuracy-precision-recall-f1-score-ade299cf63cd

Berkeley
UNIVERSITY OF CALIFORNIA

References

Images

- https://www.voxco.com/blog/employee-turnover-a-guide/
- https://en.wikipedia.org/wiki/Kaggle
- https://commons.wikimedia.org/wiki/File:Performance-Evaluation-Process-z.jpg
- https://freesvg.org/colorful-gague
- https://www.ibm.com/brand/experience-guides/developer/b1db1ae501d522a1a4b49613fe07c9f1/01_8-bar-positive.svg
- https://www.chanakya-research.com/mixed-methodology-an-alternative-research-approach-to-explore/
- Confusion Matrix adapted from Python Machine Learning 3rd Edition by Sebastian Raschka
- https://github.com/topics/scikit-learn
- https://www.shrm.org/hr-today/news/all-things-work/pages/to-have-and-to-hold.aspx

Berkeley
UNIVERSITY OF CALIFORNIA