Employee-Attrition

To Predict Employee Attrition on Fictional Dataset from IBM Data Scientists

Using data from a fictional dataset provided by IBM data scientists, four supervised machine learning algorithms in Python were used to train the models and predict employee attrition.

After fitting the models, 4 fictitious characters (Alan, Ben, Chloe, Denise) were created and their probabilities of attrition were predicted using the 4 models.

Machine Learning Algorithms

The 4 machine learning algorithms used were:

1. ***Logistic Regression***
2. ***K-Nearest Neighbors*** – A value of k=7 nearest neighbors was determined to be the optimal k value
3. ***Decision Tree*** – Original max_depth = 18 was pruned to max_depth =12
4. ***Random Forest*** – n_estimators = 2000 decision trees were used

Insights

Results from the 4 models are given below.

	Logistic Regression		KNN		Decision Tree		Random Forest	
Score on Training Data	0.89		0.86		0.99		1.0	
Score on Test Data	0.87		0.86		0.83		0.87	
Alan	**1**		**0**		**1**		**1**	
	0.42	0.58	0.71	0.29	0.0	1.0	0.41	0.59
Ben	**1**		**0**		**1**		**0**	
	0.25	0.75	0.71	0.29	0.0	1.0	0.59	0.41
Chloe	**0**		**0**		**0**		**0**	
	0.99	0.01	0.86	0.14	1.0	0.0	0.75	0.25
Denise	**0**		**0**		**0**		**0**	
	0.99	0.01	1.0	0.0	1.0	0.0	0.90	0.09

Note #1 :: **1** is Leave and **0** is Stay

Note #2 :: Probabilities for **0** (Stay) and **1** (Leave) are also appended.

A further check into the data revealed that it is heavily skewed towards employees who stayed with the company, resulting in more biased predictions, especially for KNN and Random Forest.

No. of employees who left : 237
No. of employees who stayed : 1,233
Total employees : 1,470

Looking across the probabilities table, we can conclude that

- Alan and Ben are pre-disposed to leave
- Chloe and Denise are predicted to stay

Annex : Direction of Impact of Variable on Attrition (Logistic Regression)

Variable	Direction of Impact
Age	-ve
BusinessTravel	+ve
Department_Sales	+ve
Department_RD	-ve
DistanceFromHome	+ve
Education	+ve
EducationField	-ve
EnvironmentSatisfaction	-ve
Gender	+ve
JobInvolvement	-ve
JobLevel	+ve
JobRole	+ve
JobSatisfaction	-ve
MaritalStatus	+ve

Variable	Direction of Impact
MonthlyIncome	-ve
NumCompaniesWorked	+ve
OverTime	+ve
PercentSalaryHike	-ve
PerformanceRating	-ve
RelationshipSatisfaction	-ve
StockOptionLevel	-ve
TotalWorkingYears	-ve
TrainingTimesLastYear	-ve
WorkLifeBalance	-ve
YearsAtCompany	+ve
YearsInCurrentRole	-ve
YearsSinceLastPromotion	+ve
YearsWithCurrManager	-ve

Data Source

Data was retrieved from Kaggle : https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset

Data on 1470 employees were provided, along with 34 variable features and 1 outcome label. Out of 34 features provided, 27 features were eventually selected to be included in the machine learning algorithms. 1 feature, Department, was split into two sub-features, Department_Sales and Department_RD. As such, 28 column features can be found in the models.