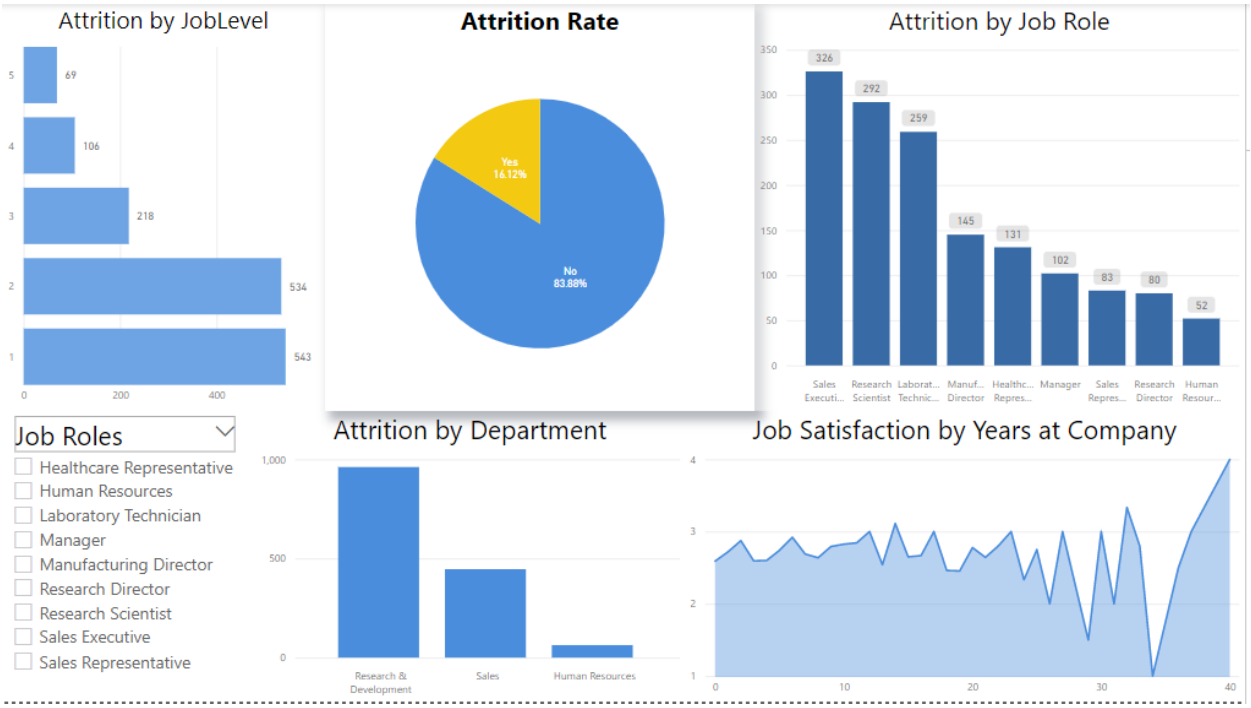
Employee Attrition Prediction

Overview

ABC company is facing attrition in their company, and they have provided data for figuring out the factors and providing the prediction on the future employees who might go through attrition. Objective is to understand the factors leading to attrition. And where the attrition in their organization is happening the most. And after knowing the general idea, with the given data we must predict who will go through the attrition in advance, so that company can deal with the issue and take required precautions.

Underlying report shows some descriptive analytics of attrition going on in their organization.



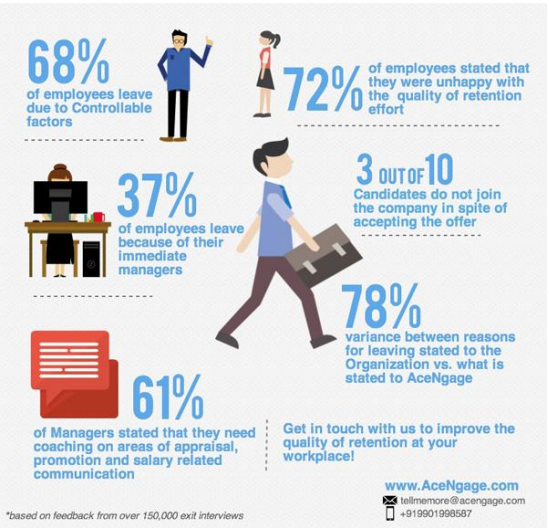
**For more interactivity please access the report in power BI*

Literature Survey

Employee attrition refers to the deliberate downsizing of a company's workforce. Downsizing happens when employees resign or retire. This type of reduction in staff is called a hiring freeze. It is one way a company can decrease labor costs without the disruption of layoffs.

There are several reasons why employee attrition takes place. They include:

- Unsatisfactory pay and/or benefits
- Lack of opportunity
- Poor workplace conditions
- Poor work-life balance
- Illness and death
- Retirement
- Relocation



The loss of employees can be a problem for corporations because it can mean the reduction of valued talent in the workforce. However, it can also be a good thing. Attrition can force a firm to identify the issues that may be causing it. It also allows companies to cut down labor costs as employees leave by choice and they're not replaced. Eventually, it can lead to the hiring of new employees with fresh ideas and energy.

Rate of attrition is described as a percentage of your total workforce. By tracking attrition rates from one year to the next, you can identify patterns and pinpoint high or low points in employee retention.

ATTRITION RATE (%) = (Number of leaves ÷ number of employees) x 100

Attrition is closely related to employee turnover or churn rate. However, turnover has more to do with the culture that you create within your company and the frequency of hiring and quitting/firing. Attrition is more concerned with the big picture and the numerical change in your workforce.

Solution

Theoretical Overview

After understanding the data, we must choose a proper methodology to perform prediction on the data. For that we are going to use a programming paradigm known as **Machine Learning**.

Machine learning (ML) is the process of using mathematical models of data to help a computer learn without direct instruction. It's considered a subset of artificial intelligence (AI). Machine learning uses algorithms to identify patterns within data, and those patterns are then used to create a data model that can make predictions. With increased data and experience, the results of machine learning are more accurate—much like how humans improve with more practice.

Machine learning techniques –

- **Supervised learning** - Addressing datasets with labels or structure, data acts as a teacher and "trains" the machine, increasing in its ability to make a prediction or decision.
- **Unsupervised learning** - Addressing datasets without any labels or structure, finding patterns and relationships by grouping data into clusters.
- **Reinforcement learning** - Replacing the human operator, an agent—a computer program acting on behalf of someone or something—helps determine outcome based upon a feedback loop.

In our case we are going to be performing **Supervised Learning**, on the data as we have both features and label for our operations.

The given problem falls under the category of **classification**. Classification is a predictive modeling problem where the class label is anticipated for a specific example of input data. For example, in determining handwriting characters, identifying spam, and so on, classification requires training data with many datasets of inputs and outputs.

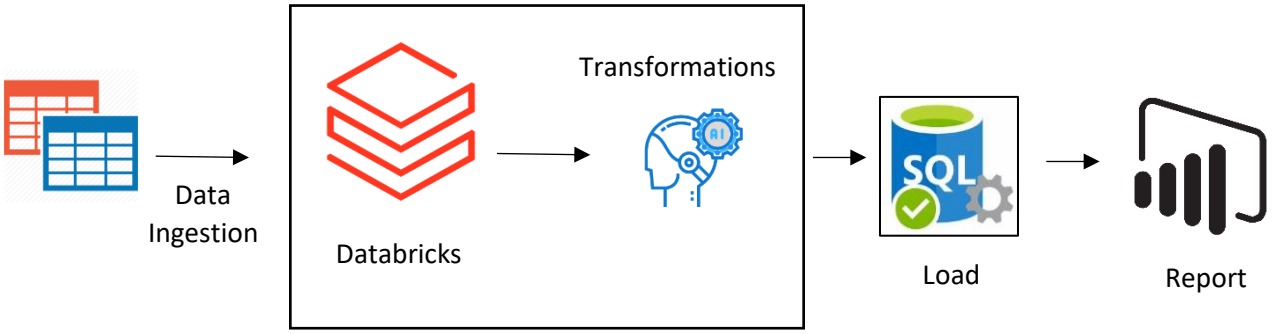
There are many classification algorithms which can be used in this, but **decision tree** with the depth 2, is used here as it is a moderate type of method, meaning it won't be over fitting for the problem or underfitting for the problem. But more models can be used also, like SVM, XGboost, neural networks etc.

Decision Tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.

A tree can be "learned" by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion is completed when the subset at a node all has the same value of the target variable, or when splitting no longer adds value to the predictions. The construction of a decision tree classifier does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery. Decision trees can handle high-dimensional data. In general decision tree classifier has good accuracy. Decision tree induction is a typical inductive approach to learn knowledge on classification.

And then a report is built with a tool, for providing overview of the result.

Technical Overview



- **Dataset**
  The key to success in any organization is attracting and retaining top talent. As an HR analyst one of the key tasks is to determine which factors keep employees at the company and which prompt others to leave. Given in the data is a set of data points on the employees who are either currently working within the company or have resigned. The objective is to identify and improve these factors to prevent loss of good people.
  The dataset consists of 1471 observations and 35 variables. Each row in dataset represents an employee; each column contains employee attributes:
  Independent Variables are:
  Age: Age of employees,
  Department: Department of work,
  Distance from home,
  Education: 1-Below College; 2-College; 3-Bachelor; 4-Master; 5-Doctor.
  Education Field
  Environment Satisfaction: 1-Low; 2-Medium; 3-High; 4-Very High.
  Job Satisfaction: 1-Low; 2-Medium; 3-High; 4-Very High.
  Marital Status,
  Monthly Income,
  Job level,
  Job role,
  Num Companies Worked: Number of companies worked prior to IBM,
  Work Life Balance: 1-Bad; 2-Good; 3-Better; 4-Best.
  Years At Company: Current years of service in IBM
  Dependent Variable was:
  Attrition: Employee attrition status (0 or 1)
  And many more.

**A detailed description of machine learning process is provided in notebook.*

Tools description –

- **Databricks -** The Azure Databricks Lakehouse Platform provides a unified set of tools for building, deploying, sharing, and maintaining enterprise-grade data solutions at scale. Azure Databricks integrates with cloud storage and security in your cloud account and manages and deploys cloud infrastructure on your behalf.
    - **Libraries used –**
      NumPy
      Pandas
      Seaborn
      Scikit learn
- **Power BI -** Power BI is a unified, scalable platform for self-service and enterprise business intelligence (BI). Connect to and visualize any data, and seamlessly infuse the visuals into the apps you use every day.
- **Azure SQL -** Azure SQL Database is a fully managed platform as a service (PaaS) database engine that handles most of the database management functions such as upgrading, patching, backups, and monitoring without user involvement.

Solution Implementation Overview

- Data is ingested as Pyspark data frame. And then some transformations like removing few columns which are not stating factual meanings, are removed.

```
df_main = spark.read.csv("dbfs:/FileStore/WA_Fn_UseC__HR_Employee_Attrition.csv", header=True, inferSchema=True)
```

```
df_main.display()
```

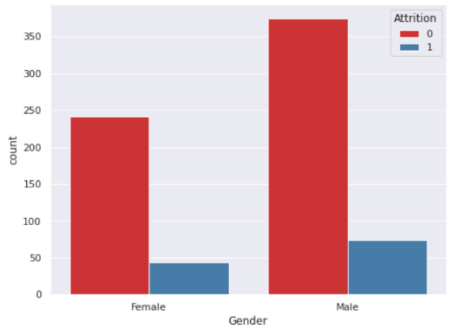
Table

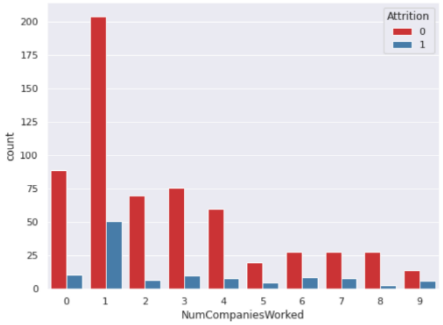
	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	Emp
1	41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences	1	1
2	49	No	Travel_Frequently	279	Research & Development	8	1	Life Sciences	1	2
3	37	Yes	Travel_Rarely	1373	Research & Development	2	2	Other	1	4
4	33	No	Travel_Frequently	1392	Research & Development	3	4	Life Sciences	1	5
5	27	No	Travel_Rarely	591	Research & Development	2	1	Medical	1	7
6	32	No	Travel_Frequently	1005	Research & Development	2	2	Life Sciences	1	8
7	59	No	Travel_Rarely	1324	Research & Development	3	3	Medical	1	10

1,000 rows | Truncated data

- Data is prepared for machine learning process, for that data is went through preprocessing stage.

  Data preprocessing can refer to manipulation or dropping of data before it is used to ensure or enhance performance and is an important step in the data mining process. The phrase "garbage in, garbage out" is particularly applicable to data mining and machine learning projects. Data-gathering methods are often loosely controlled, resulting in out-of-range values (e.g., Income: −100), impossible data combinations (e.g., Sex: Male, Pregnant: Yes), and missing values, etc.
- And then relationship between every feature is analyzed with the help of Seaborn.
  For example,





- Now the data is split in train and test.

```
X = new_raw_data.drop('Attrition', axis=1).values# Input features (attributes)
y = new_raw_data['Attrition'].values # Target vector
print('X shape: {}'.format(np.shape(X)))
print('y shape: {}'.format(np.shape(y)))

X_train, X_test, y_train, y_test = train_test_split(X, y, train_size = 0.7, test_size=0.3, random_state=0)

X shape: (735, 15)
y shape: (735,)
```

- Then decision tree is implemented.

```
dt = DecisionTreeClassifier(criterion='entropy', max_depth=2, random_state=1)
dt.fit(X_train, y_train)

Out[77]: DecisionTreeClassifier(criterion='entropy', max_depth=2, random_state=1)
```

- And after training the model, major features affecting the decision are printed.

	index	Variable	Feature Importance Score
0	4	TotalWorkingYears	0.622185
1	0	Age	0.287432
2	8	YearsSinceLastPromotion	0.090382
3	1	MonthlyIncome	0.000000
4	2	NumCompaniesWorked	0.000000
5	3	PercentSalaryHike	0.000000
6	5	TrainingTimesLastYear	0.000000
7	6	YearsAtCompany	0.000000
8	7	YearsInCurrentRole	0.000000
9	9	YearsWithCurrManager	0.000000
10	10	MaritalStatus_Divorced	0.000000

- And confusion matrix is printed

- And then model is deployed to work on unseen data and then output is appended in the main data and then sent to the sql server provisioned on azure and from there it is connected to power bi for making a report.

	Age	BusinessTravel	Department	DistanceFromHome	EducationField	EnvironmentSatisfaction	Gender	JobLevel	JobRole	JobSatisfaction	...	RelationshipSatisfaction	TotalWor
0	41	Travel_Rarely	Sales	1	Life Sciences	2	Female	2	Sales Executive	4	...	1	
1	49	Travel_Frequently	Research & Development	8	Life Sciences	3	Male	2	Research Scientist	2	...	4	
2	37	Travel_Rarely	Research & Development	2	Other	4	Male	1	Laboratory Technician	3	...	2	
3	33	Travel_Frequently	Research & Development	3	Life Sciences	4	Female	1	Research Scientist	3	...	3	
4	27	Travel_Rarely	Research & Development	2	Medical	1	Male	1	Laboratory Technician	2	...	4	

5 rows × 24 columns

Connecting to SQL

```
jdbcHostname = "databaseyr.database.windows.net"
jdbcPort = 1433
jdbcDatabase = "db_training"
jdbcUsername = "trianee_YR"
jdbcPassword = "Pa$$word1234"

jdbcUrl = f"jdbc:sqlserver://{jdbcHostname}:{jdbcPo
```

```
output_df.write.format("jdbc") \
  .mode("overwrite") \
  .option("url", jdbcUrl) \
  .option("dbtable", "attrition_output") \
  .option("user", jdbcUsername) \
  .option("password", jdbcPassword) \
  .save()
```