Employee Churn Analysis

USING MACHINE LEARNING

DMPA LAB PROJECT

- Analysis of dataset

Overview

Overview | Reproduction | Warnings 1

Dataset statistics

Number of variables	10
Number of observations	14999
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	3008
Duplicate rows (%)	20.1%
Total size in memory	1.1 MiB
Average record size in memory	80.0 B

Variable types

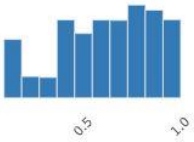
NUM	5
BOOL	3
CAT	2

Variables

satisfaction_level
Real number ($\mathbb{R}_{\geq 0}$)

Distinct count	92
Unique (%)	0.6%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%

Mean	0.6128335222348156
Minimum	0.09
Maximum	1.0
Zeros	0
Zeros (%)	0.0%
Memory size	117.2 KiB

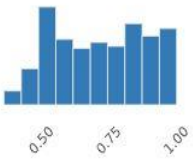


Toggle details

last_evaluation
Real number ($\mathbb{R}_{\geq 0}$)

Distinct count	65
Unique (%)	0.4%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%

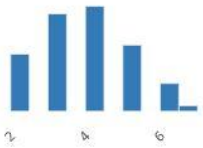
Mean	0.7161017401160078
Minimum	0.36
Maximum	1.0
Zeros	0
Zeros (%)	0.0%
Memory size	117.2 KiB



Toggle details

number_project
Real number ($\mathbb{R}_{\geq 0}$)

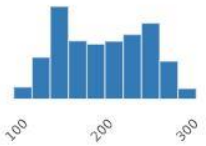
Distinct count	6	**Mean**	3.80305353690246
Unique (%)	< 0.1%	**Minimum**	2
Missing	0	**Maximum**	7
Missing (%)	0.0%	**Zeros**	0
Infinite	0	**Zeros (%)**	0.0%
Infinite (%)	0.0%	**Memory size**	117.2 KiB



Toggle details

average_montly_hours
Real number ($\mathbb{R}_{\geq 0}$)

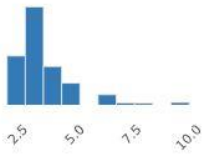
Distinct count	215	**Mean**	201.0503366891126
Unique (%)	1.4%	**Minimum**	96
Missing	0	**Maximum**	310
Missing (%)	0.0%	**Zeros**	0
Infinite	0	**Zeros (%)**	0.0%
Infinite (%)	0.0%	**Memory size**	117.2 KiB



Toggle details

time_spend_company
Real number ($\mathbb{R}_{\geq 0}$)

Distinct count	8	**Mean**	3.498233215547703
Unique (%)	0.1%	**Minimum**	2
Missing	0	**Maximum**	10
Missing (%)	0.0%	**Zeros**	0
Infinite	0	**Zeros (%)**	0.0%
Infinite (%)	0.0%	**Memory size**	117.2 KiB



Toggle details

Work_accident
Boolean

Distinct count	2
Unique (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	117.2 KiB

Toggle details

promotion_last_5years
Boolean

Distinct count	2
Unique (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	117.2 KiB

Toggle details

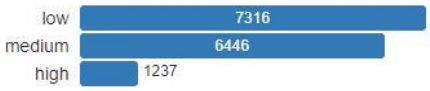
quit
Boolean

Distinct count	2
Unique (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	117.2 KiB

Toggle details

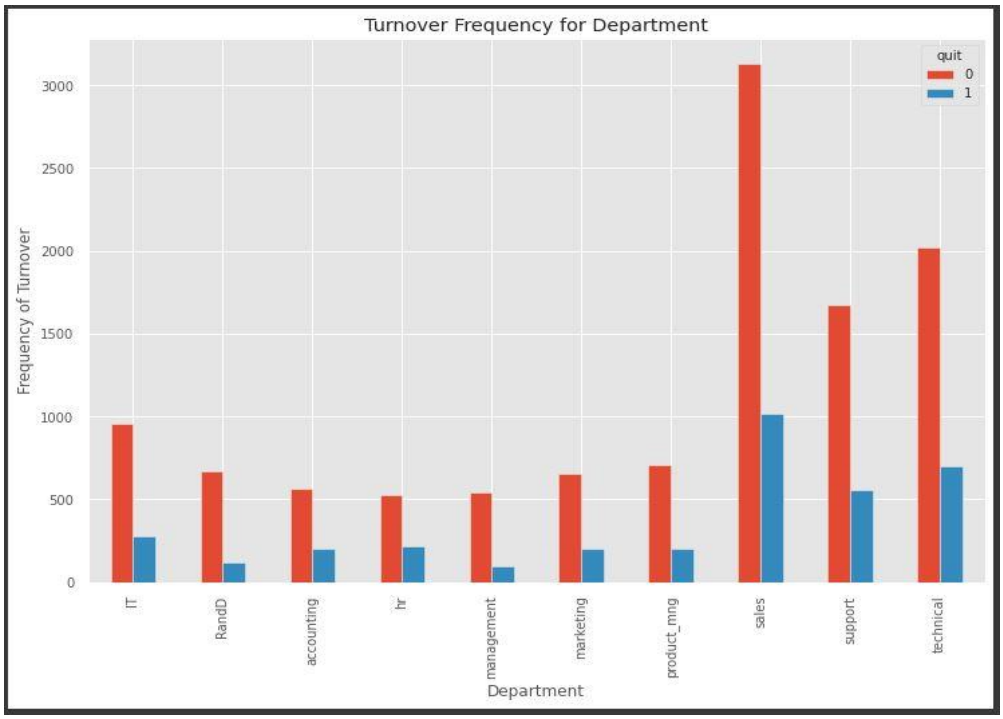
salary
Categorical

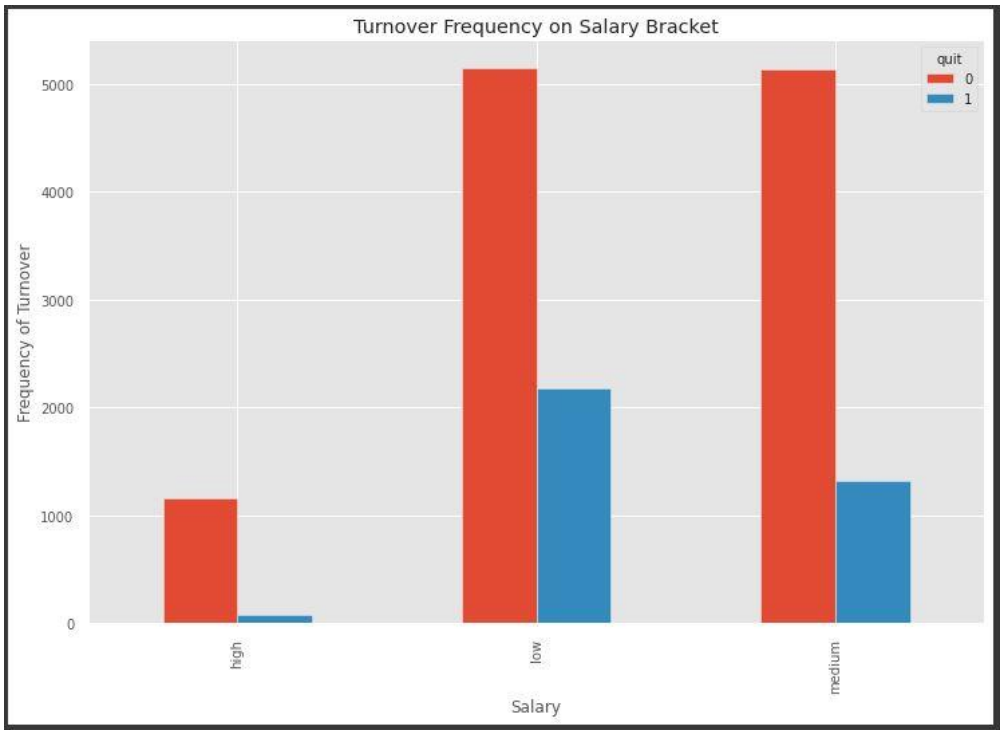
Distinct count	3
Unique (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	117.2 KiB



Toggle details

- ## ANALYSIS OF CATEGORICAL FEATURES





- ## COMPARISON OF ALGORITHMS USED

1. Decision Tree

```
from sklearn.tree import DecisionTreeClassifier
from sklearn import metrics

# Training the Decision Tree Classification model on the Training set
model=DecisionTreeClassifier()
model=model.fit(X_train,y_train)

# Prediction
y_pred=model.predict(X_test)

# Accuracy
print("\nAccuracy (Decision Tree Classifier Model):",metrics.accuracy_score(y_test, y_pred))
```

```
Accuracy (Decision Tree Classifier Model): 0.981
```

crit gini

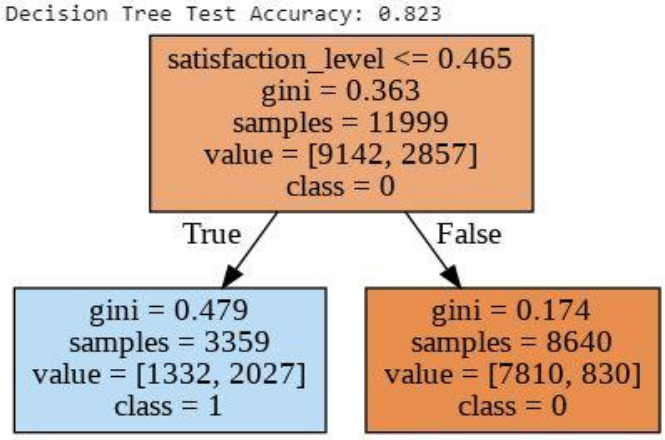
split best

depth 1

min_split 2

min_leaf 1

```
Decision Tree Training Accuracy: 0.820
Decision Tree Test Accuracy: 0.823
```




```
DecisionTreeClassifier(ccp_alpha=0.0, class_weight=None, criterion='gini',
                       max_depth=1, max_features=None, max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, presort='deprecated',
                       random_state=0, splitter='best')
```

Widget for analysis of Decision Tree

crit gini

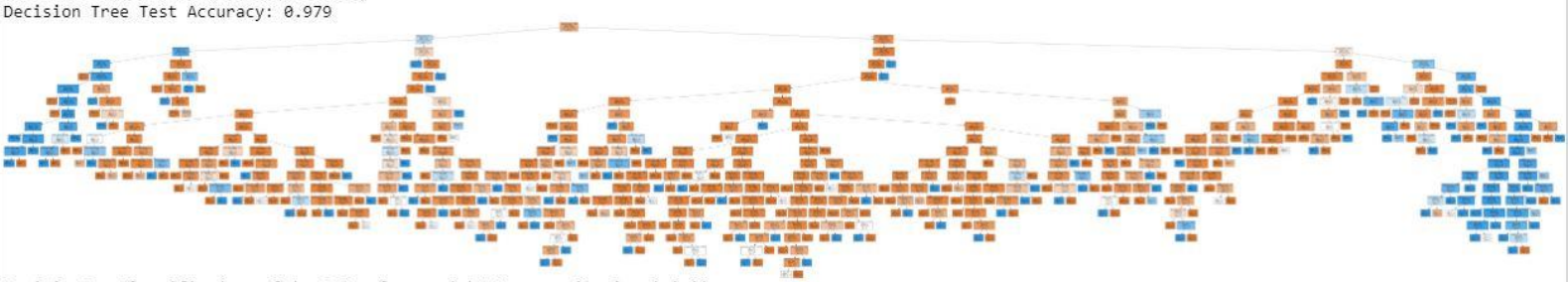
split best

depth 20

min_split 4

min_leaf 1

```
Decision Tree Training Accuracy: 0.997
Decision Tree Test Accuracy: 0.979
```

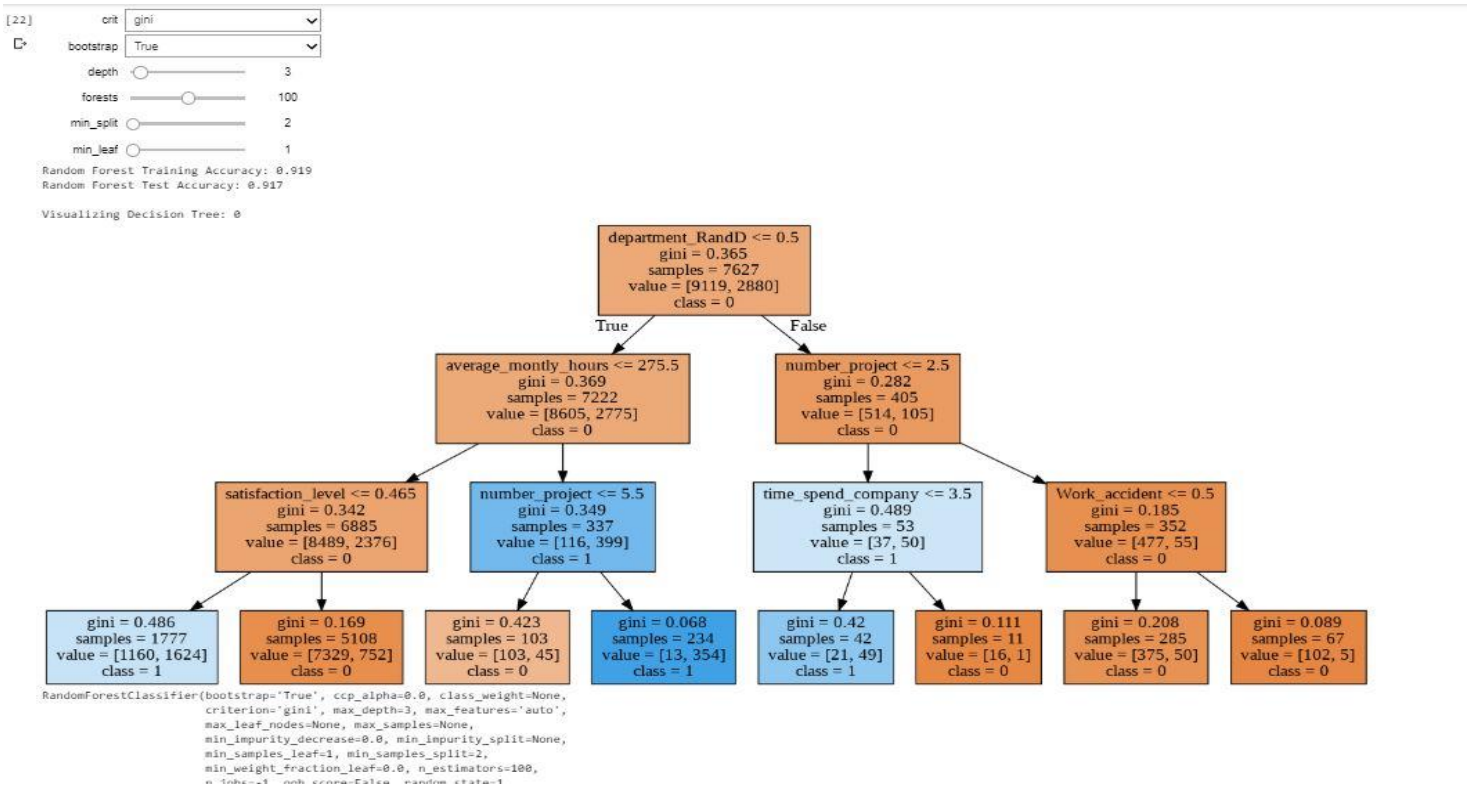


```
DecisionTreeClassifier(ccp_alpha=0.0, class_weight=None, criterion='gini',
                       max_depth=30, max_features=None, max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, presort='deprecated',
                       random_state=0, splitter='best')
DecisionTreeClassifier(ccp_alpha=0.0, class_weight=None, criterion='gini',
                       max_depth=20, max_features=None, max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=4,
                       min_weight_fraction_leaf=0.0, presort='deprecated',
                       random_state=0, splitter='best')
```

2. Random Forest

```
[21] # Training the Random Forest Classification model on the Training set

     from sklearn.ensemble import RandomForestClassifier
     classifier = RandomForestClassifier(n_estimators = 10, criterion = 'entropy', random_state = 0)
     classifier.fit(X_train, y_train)

     # Prediction
     y_pred = classifier.predict(X_test)

     # Accuracy
     print("\nAccuracy (Random Forest Classifier Model):",metrics.accuracy_score(y_test, y_pred))
```

```
Accuracy (Random Forest Classifier Model): 0.988
```



Widget for analysis of Random Forest

3. KNN

```
[26] # Training the K-NN model on the Training set
     from sklearn.neighbors import KNeighborsClassifier
     classifier = KNeighborsClassifier(n_neighbors = 5, metric = 'minkowski', p = 2)
     classifier.fit(X_train, y_train)

     # Prediction
     y_pred = classifier.predict(X_test)

     # Accuracy
     print("\nAccuracy (Random Forest Classifier Model):",metrics.accuracy_score(y_test, y_pred))
```

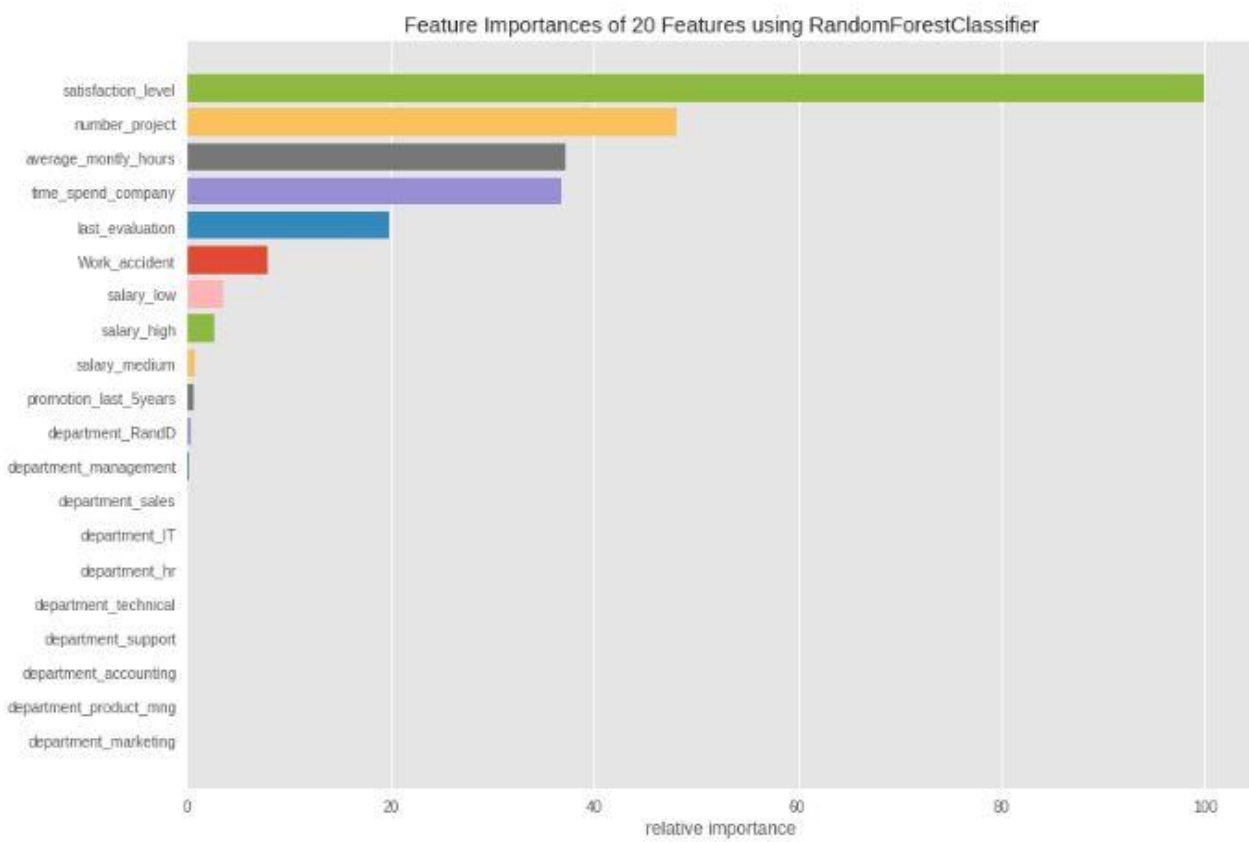
```
Accuracy (Random Forest Classifier Model): 0.9326666666666666
```

4. Logistic Regression

```
[24] from sklearn.linear_model import LogisticRegressionCV

     logit = LogisticRegressionCV(random_state=1, n_jobs=-1,max_iter=500,
                                  cv=10)

     lr = logit.fit(X_train, y_train)

     print('Logistic Regression Accuracy: {:.3f}'.format(accuracy_score(y_test, lr.predict(X_test))))
```

```
Logistic Regression Accuracy: 0.790
```

- Relative Importance of Features



For complete analysis, graphs and code visit the following link:

https://github.com/apoorv-mit2021/Employee-Churn-Analysiss