Котолевский Максим Николаевич, группа 19

Лабораторная работа № 3

Вариант № 5

Цель работы

Целью данной работы является прогнозирование оттока сотрудников из компании и выбор самых важных показателей

Задание

Решить задачу предсказания оттока сотрудников.

Код программы (внесённые изменения в шаблон кода выделены)

Ссылка на исходный dataset:
https://www.kaggle.com/datasets/liujiaqi/hr-comma-sepcsv

Код программы (внесённые изменения в шаблон кода выделены)

```
import pandas as pd
from sklearn import preprocessing
from sklearn.model_selection import train_test_split
from sklearn.ensemble import GradientBoostingClassifier,
RandomForestRegressor
from sklearn import metrics
```

```
data=pd.read_csv('HR_comma_sep.csv')
```

```
data.shape
```

```
data.info()
```

```
data.describe()
```

```
#Так как модель работает только с числовыми признаками, нужно привести все
категориальные признаки к числовым

le = preprocessing.LabelEncoder()

data['salary']=le.fit_transform(data['salary'])
data['sales']=le.fit_transform(data['sales'])
```

```
#Разделим данные на тестовые и обучающие
X=data.drop('left', axis=1)
y=data['left']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,
random_state=42)  #70%/30%
```

```
#Используем градиентный бустинг для обучения модели
gb = GradientBoostingClassifier()
gb.fit(X_train, y_train)
y_pred = gb.predict(X_test)
```

```
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

```
#Accuracy: 0.9715555555555555
```

```
#Выделяем самые важные признаки
gb.feature_importances_
```

```
feature_importance = pd.DataFrame(columns = ['Название', 'Важность'])
```

```
feature_importance['Название'] = X_test.keys()
feature_importance['Важность'] = gb.feature_importances_
```

```
feature_importance.sort_values(by='Важность', ascending=False)
```

Результаты выполнения задания

Было разработано ПО для реализации задачи прогнозирования оттока сотрудников из компании и выбора самых важных показателей.

Out[31]:

	Название	Важность
0	satisfaction_level	0.553191
4	time_spend_company	0.169690
2	number_project	0.105444
1	last_evaluation	0.100397
3	average_montly_hours	0.068911
5	Work_accident	0.001045
8	salary	0.001008
7	sales	0.000315
6	promotion_last_5years	0.000000

Точность с использованием градиентного бустинга оказалась 97%. Наиболее важным оказался показатель уровня удовлетворенности.