Основы машинного обучения. Python для анализа данных

Анализ маркетинговых показателей, решение задач регрессии

Презентацию подготовила
Загирова Анастасия

Описание

Задача: Провести полноценный анализ данных источника, используя Python и соответствующие библиотеки для анализа данных. Решить задачи регрессии или классификации

Для этого необходимо:

- Провести исследовательский анализ данных;
- Сформулировать гипотезы для машинного обучения;
- Провести разведочный анализ данных;
- Проверить гипотезы роазличными методами машинного обучения;
- Выбрать самый эффективный метод обучения.

Источники данных: Анализ маркетинговых показателей:https://www.kaggle.com/jackdaoud/marketing-data

Исследовательский анализ данных

- Столбцы переименованы, удалены пробелы в названиях, приведен к нижнему регистру
- Пропуски не обнаружены
- Выполнена проверка на дубликаты. Процент дубликатов 1.77
- Вычислена дисперсия, значения моды

Датасет имеет 39 колонок и 2205 строк

Гипотеза №1

Спрогнозировать общую сумму трат для нового покупателя.

Целевая переменная - mnt_total - общая сумма трат

Признаки - 14 столбцов. 4 столбца с числовыми данными, 10 столбцов с категориальными данными

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2205 entries, 0 to 2204
Data columns (total 15 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   income                2205 non-null   float64
 1   kids                  2205 non-null   int64
 2   teen                  2205 non-null   int64
 3   age                   2205 non-null   int64
 4   marital_divorced      2205 non-null   int64
 5   marital_married       2205 non-null   int64
 6   marital_single        2205 non-null   int64
 7   marital_together      2205 non-null   int64
 8   marital_widow         2205 non-null   int64
 9   education_2ncycle     2205 non-null   int64
 10  education_basic       2205 non-null   int64
 11  education_graduation  2205 non-null   int64
 12  education_master      2205 non-null   int64
 13  education_phd         2205 non-null   int64
 14  mnt_total             2205 non-null   int64
dtypes: float64(1), int64(14)
memory usage: 258.5 KB
```

Линейная регрессия. Метод лассо

Результаты обучения

```
Линейная регрессия
R2 train: 0.7206255172508218
R2 test: 0.7653645052636835
Train MSE: 304.10986390599754
Test MSE: 279.77421259226594

Лассо - регрессия
R2 train: 0.7203840964888549
R2 test: 0.7646165728708054
Train MSE: 304.24123342139427
Test MSE: 280.21976689096624
```

Среднее значение целевой переменной 558

MSE тестовых данных методом лассо немного хуже, поскольку больше. Соответственно считаем линейную регрессию более эффективной моделью.

Гипотеза №2

Спрогнозируем вероятность того, что покупатель оставит отзыв

Целевая переменная - response - категориальный столбец 1- есть отзыв, 0 - нет отзыва

Признаки - 14 столбцов. 8 столбцов с числовыми данными, 6 столбцов с категориальными данными

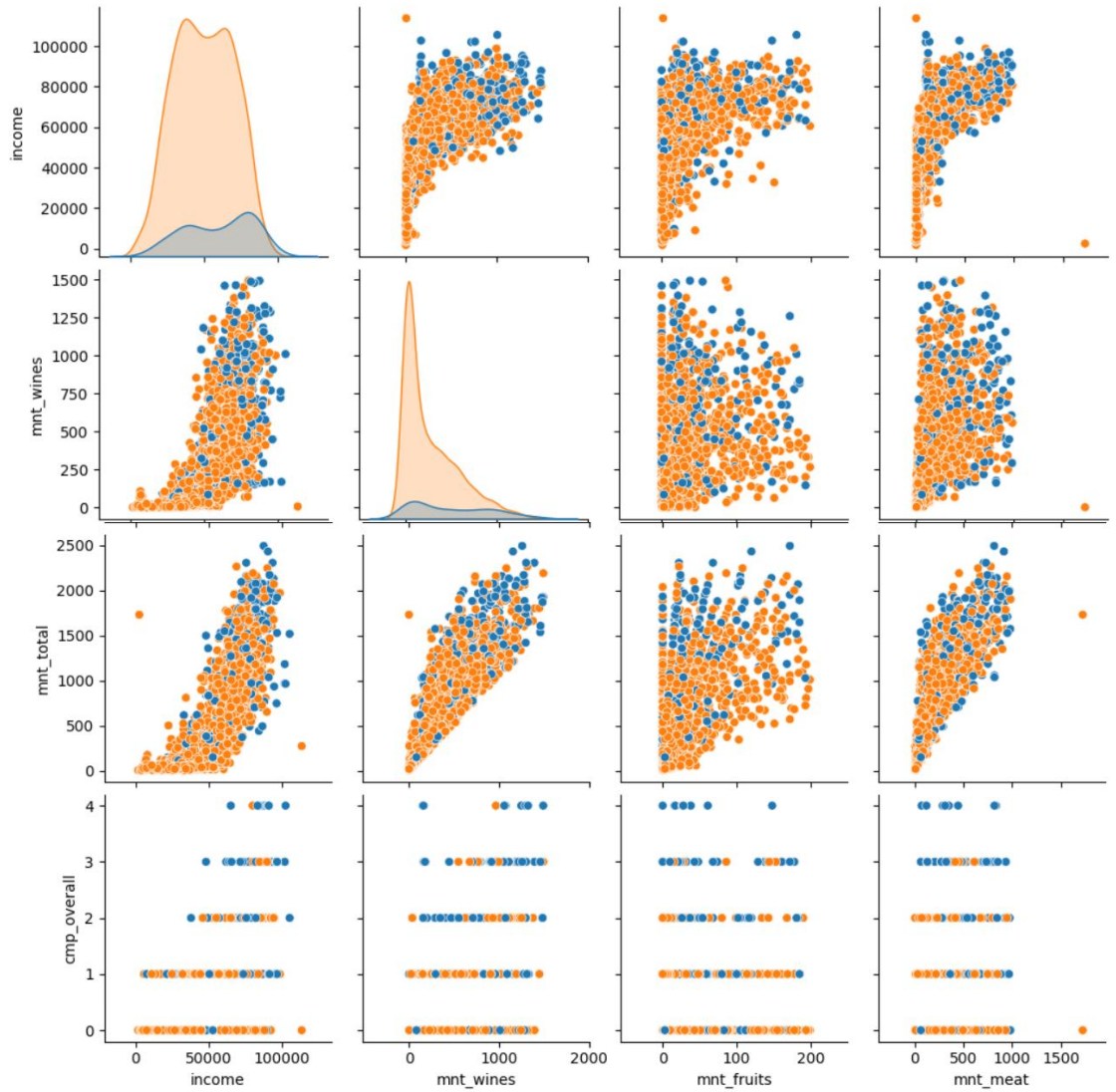
Выборки несбалансированы

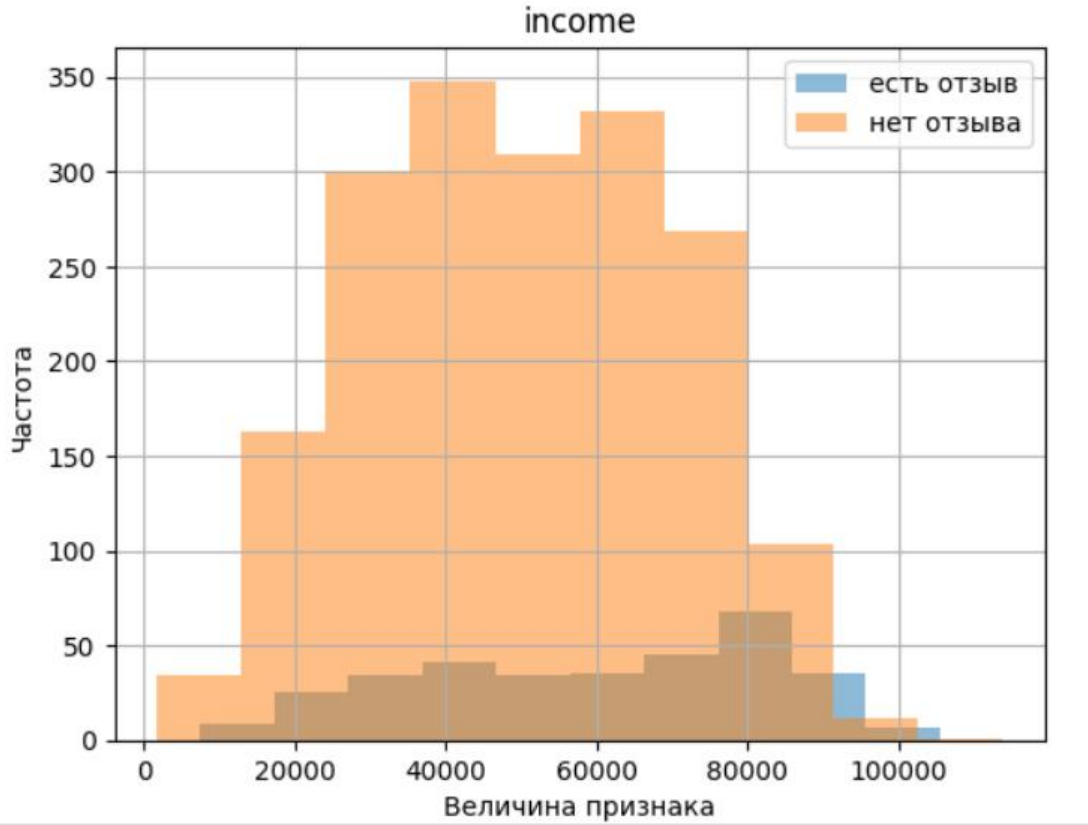
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2205 entries, 0 to 2204
Data columns (total 15 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   income       2205 non-null   float64
 1   mnt_wines    2205 non-null   int64
 2   mnt_fruits   2205 non-null   int64
 3   mnt_meat     2205 non-null   int64
 4   mnt_fish     2205 non-null   int64
 5   mnt_sweet    2205 non-null   int64
 6   mnt_gold     2205 non-null   int64
 7   cmp3         2205 non-null   int64
 8   cmp4         2205 non-null   int64
 9   cmp5         2205 non-null   int64
 10  cmp1         2205 non-null   int64
 11  cmp2         2205 non-null   int64
 12  response     2205 non-null   int64
 13  mnt_total    2205 non-null   int64
 14  cmp_overall  2205 non-null   int64
dtypes: float64(1), int64(14)
memory usage: 258.5 KB
```

```
Баланс выборок:
 0    1872
1     333
Name: response, dtype: int64
```

Распределение признаков





Построены парные диаграммы рассеяния, гисторгаммы признаков

Логистической регрессия и решающее дерево

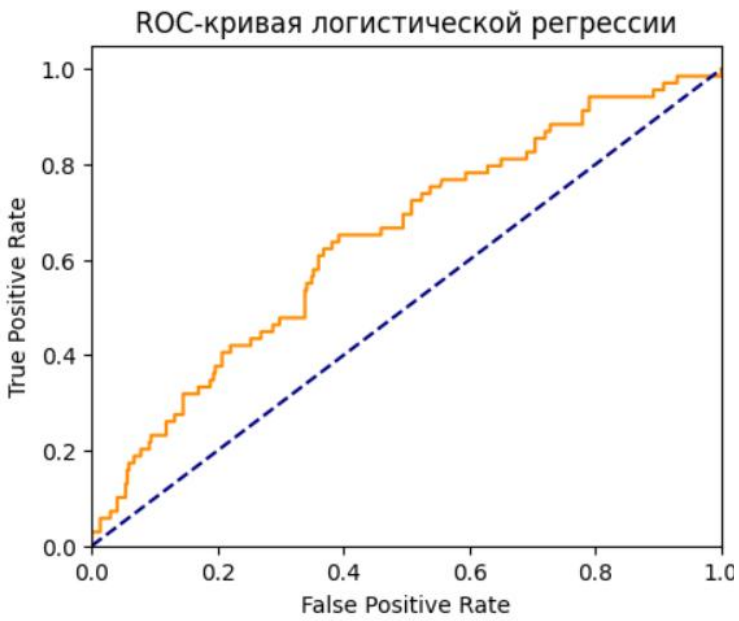
Результаты обучения

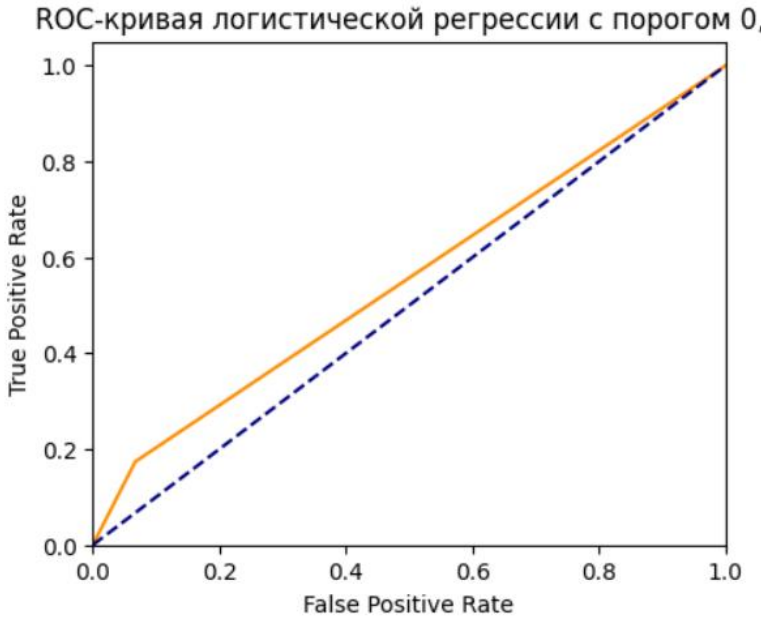
```
Логистическая регрессия
accuracy:  0.83
f1:  0.15
rocauc:  0.64

Логистическая регрессия с измененным порогом
accuracy:  0.81
f1:  0.23
rocauc:  0.55

Решающее дерево
accuracy:  0.8
f1:  0.38
rocauc:  0.64
```

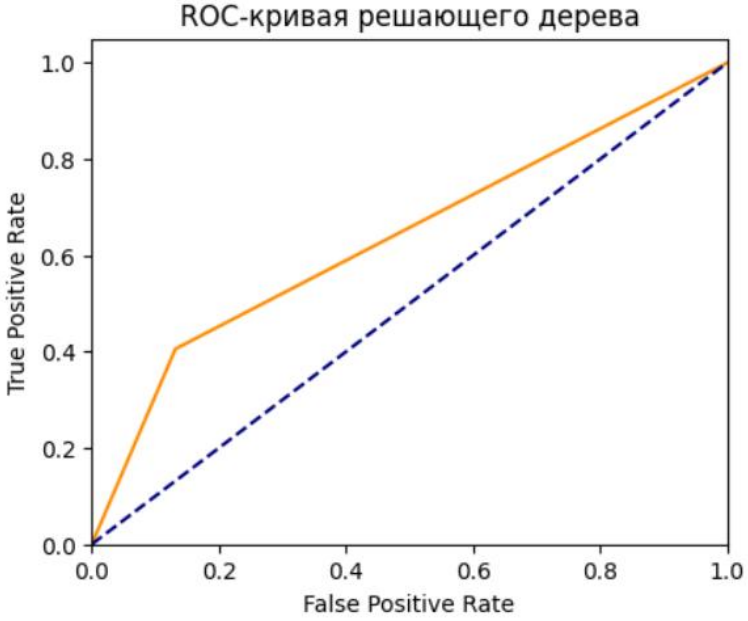
0,4

В данном случае дерево решений оказалось самым эффективным методом. При почти такой же accuracy, такой же rocauc, как и у логистической регрессии, у дерева решений значительно выше метрика f1.

ROC-кривая







В данном случае дерево решений оказалось самым эффективным методом. При почти такой же accuracy, такой же rocauc, как и у логистической регрессии, у дерева решений значительно выше метрика f1.

Выводы

Мы проанализировали данные маркетинговых метриках.
На основании имеющихся данных было исследовано 2 гипотезы машинного обучения с помощью 4 методов.

- линейная регрессия,
- метод лассо,
- логистическая регрессия,
- решающее дерево

Основываясь на показаниях ключевых метрик оценки качества моделей mse, R2, accurancy, f1, rocauc можно сказать, что все модели хорошо показывают себя в прогнозировании.

В первом случае более эффективной показала себя модель **линейной регрессии**. Для второй гипотезы будем считать более эффективной модель **решающего дерева**.