МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение высшего образования

НИЖЕГОРОДСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ

УНИВЕРСИТЕТ им. Р.Е.АЛЕКСЕЕВА

Институт радиоэлектроники и информационных технологий

Кафедра информатики и систем управления

ОТЧЕТ

по лабораторной работе №3

по дисциплине

Предиктивная аналитика

РУКОВОДИТЕЛЬ:

_____ Санников А.Н.

(подпись) (фамилия, и.,о.)

СТУДЕНТ:

_____ Напылов Е.И.

(подпись) (фамилия, и.,о.)

_М22-ИВТ-1_____

(шифр группы)

Работа защищена «___» _____

С оценкой _____

Нижний Новгород 2023

Содержание

1. Постановка задачи

В данной работе необходимо решить задачу линейной регрессии.

Была выбрана задача определения цены авиаперелета по некоторым признакам. Нужно построить линейную регрессию, которая будет определять стоимость полета. Датасет содержит более 10000 записей о совершенных перелетах. Всего имеется 11 признаков:

1. 'Airline' - авиакомпания
2. 'Date_of_Journey' - дата
3. 'Source' - откуда
4. 'Destination' - куда
5. 'Route' - маршрут по аэропортам
6. 'Dep_Time' - время отправления
7. 'Arrival_Time' - время прибытия
8. 'Duration' - продолжительность полета
9. 'Total_Stops' - число остановок
10. 'Additional_Info' - доп инфа
11. 'Price' - цена полета (таргет)

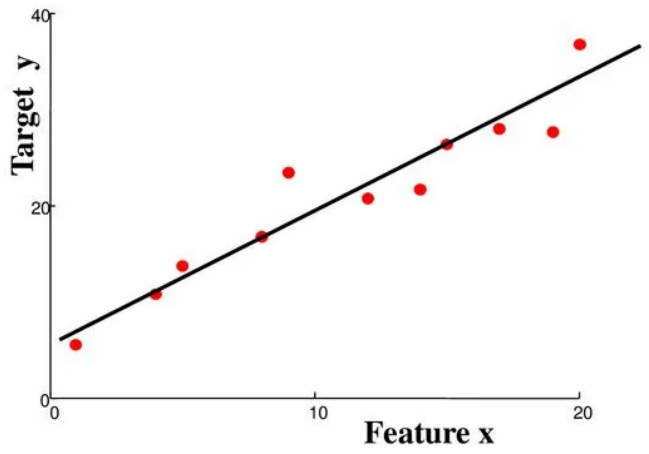
Датасет выглядит следующим образом:

	Airline	Date_of_Journey	Source	Destination	Route	Dep_Time	Arrival_Time	Duration	Total_Stops	Additional_Info	Price
0	IndiGo	24/03/2019	Banglore	New Delhi	BLR → DEL	22:20	01:10 22 Mar	2h 50m	non-stop	No info	3897
1	Air India	1/05/2019	Kolkata	Banglore	CCU → IXR → BBI → BLR	05:50	13:15	7h 25m	2 stops	No info	7662
2	Jet Airways	9/06/2019	Delhi	Cochin	DEL → LKO → BOM → COK	09:25	04:25 10 Jun	19h	2 stops	No info	13882
3	IndiGo	12/05/2019	Kolkata	Banglore	CCU → NAG → BLR	18:05	23:30	5h 25m	1 stop	No info	6218
4	IndiGo	01/03/2019	Banglore	New Delhi	BLR → NAG → DEL	16:50	21:35	4h 45m	1 stop	No info	13302
...	...	...	...	...	...	...	...	...	...	...	...
10678	Air Asia	9/04/2019	Kolkata	Banglore	CCU → BLR	19:55	22:25	2h 30m	non-stop	No info	4107
10679	Air India	27/04/2019	Kolkata	Banglore	CCU → BLR	20:45	23:20	2h 35m	non-stop	No info	4145
10680	Jet Airways	27/04/2019	Banglore	Delhi	BLR → DEL	08:20	11:20	3h	non-stop	No info	7229
10681	Vistara	01/03/2019	Banglore	New Delhi	BLR → DEL	11:30	14:10	2h 40m	non-stop	No info	12648
10682	Air India	9/05/2019	Delhi	Cochin	DEL → GOI → BOM → COK	10:55	19:15	8h 20m	2 stops	No info	11753

2. Линейная регрессия

Линейная регрессия - это один из наиболее широко используемых алгоритмов машинного обучения в задачах регрессии, когда нужно предсказать непрерывную величину. Она позволяет найти линейную зависимость между независимыми и зависимыми переменными на основе обучающих данных, после чего можно использовать полученную модель для предсказания значений зависимой переменной на новых данных.

Математически, линейная регрессия моделирует зависимость между независимой переменной X и зависимой переменной Y с помощью линейной функции: Y = aX + b, где a и b - коэффициенты модели, которые необходимо определить в процессе обучения, чтобы минимизировать ошибку предсказания.



Обычно используется метод наименьших квадратов (OLS), чтобы определить значения коэффициентов модели. Он минимизирует сумму квадратов ошибок (SSE), которые являются разностями между реальными и предсказанными значениями зависимой переменной. SSE вычисляется как: SSE = Σ(y - y*)^2, где y - реальное значение зависимой переменной, y* - предсказанное значение зависимой переменной.

Однако, линейная регрессия может также применяться к многомерным данным, где зависимая переменная зависит от нескольких независимых переменных. В этом случае, модель выглядит как: Y = a1X1 + a2X2 + ... + anXn + b, где a1, a2, ..., an - коэффициенты модели, которые необходимо определить в процессе обучения, X1, X2, ..., Xn - независимые переменные, b - свободный член.

Линейная регрессия часто используется для анализа и прогнозирования временных рядов, анализа рынка и экономики, медицинских и социальных исследований, и многих других областей, где необходимо предсказывать непрерывные значения на основе некоторых факторов.

3. Данные и их обработка

Датасет содержит более 10000 записей о перелетах и 11 признаков (включая таргет).

	Airline	Date_of_Journey	Source	Destination	Route	Dep_Time	Arrival_Time	Duration	Total_Stops	Additional_Info	Price
0	IndiGo	24/03/2019	Banglore	New Delhi	BLR → DEL	22:20	01:10 22 Mar	2h 50m	non-stop	No info	3897
1	Air India	1/05/2019	Kolkata	Banglore	CCU → IXR → BBI → BLR	05:50	13:15	7h 25m	2 stops	No info	7662
2	Jet Airways	9/06/2019	Delhi	Cochin	DEL → LKO → BOM → COK	09:25	04:25 10 Jun	19h	2 stops	No info	13882
3	IndiGo	12/05/2019	Kolkata	Banglore	CCU → NAG → BLR	18:05	23:30	5h 25m	1 stop	No info	6218
4	IndiGo	01/03/2019	Banglore	New Delhi	BLR → NAG → DEL	16:50	21:35	4h 45m	1 stop	No info	13302
...	...	...	...	...	...	...	...	...	...	...	...
10678	Air Asia	9/04/2019	Kolkata	Banglore	CCU → BLR	19:55	22:25	2h 30m	non-stop	No info	4107
10679	Air India	27/04/2019	Kolkata	Banglore	CCU → BLR	20:45	23:20	2h 35m	non-stop	No info	4145
10680	Jet Airways	27/04/2019	Banglore	Delhi	BLR → DEL	08:20	11:20	3h	non-stop	No info	7229
10681	Vistara	01/03/2019	Banglore	New Delhi	BLR → DEL	11:30	14:10	2h 40m	non-stop	No info	12648
10682	Air India	9/05/2019	Delhi	Cochin	DEL → GOI → BOM → COK	10:55	19:15	8h 20m	2 stops	No info	11753

10462 rows × 11 columns

Была проделана достаточно большая работа по обработке данных:

1. Удаление пропущенных значений
2. Стандартизация времени прибытия (иногда там была дата)
3. Преобразование времени полета в минуты
4. Преобразование даты отправления в дни с начала года (год всегда 2019)
5. Преобразование количества остановок из странных строк в нормальные числа
6. Преобразование времени отправления и прибытия в минуты (с начала суток)
7. Устранение выбросов по цене
8. OneHot encoding для пунктов отправления и назначения, авиакомпаний, дополнительной информации
9. Label encoding для маршрута (OneHot дал переобучение)

После обработки датасет выглядит так:

	Date_of_Journey	Route	Dep_Time	Arrival_Time	Duration	Total_Stops	Price	Source_Chennai	Source_Delhi	Source_Kolkata	...	Multiple carriers	Multiple carriers Premium economy	SpiceJet	Trujet	Vistara	Vistara Premium economy	Change airports	In-flight meal not included	No check-in baggage included	Red-eye flight
0	114	18	1340	70	170	0	3897	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	151	78	350	795	445	2	7662	0	0	1	...	0	0	0	0	0	0	0	0	0	0
2	189	110	565	265	1140	2	13882	0	1	0	...	0	0	0	0	0	0	0	0	0	0
3	162	85	1085	1410	325	1	6218	0	0	1	...	0	0	0	0	0	0	0	0	0	0
4	91	29	1010	1295	285	1	13302	0	0	0	...	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
10678	129	58	1195	1345	150	0	4107	0	0	1	...	0	0	0	0	0	0	0	0	0	0
10679	147	58	1245	1400	155	0	4145	0	0	1	...	0	0	0	0	0	0	0	0	0	0
10680	147	18	500	680	180	0	7229	0	0	0	...	0	0	0	0	0	0	0	0	0	0
10681	91	18	690	850	160	0	12648	0	0	0	...	0	0	0	0	1	0	0	0	0	0
10682	159	101	655	1155	500	2	11753	0	1	0	...	0	0	0	0	0	0	0	0	0	0

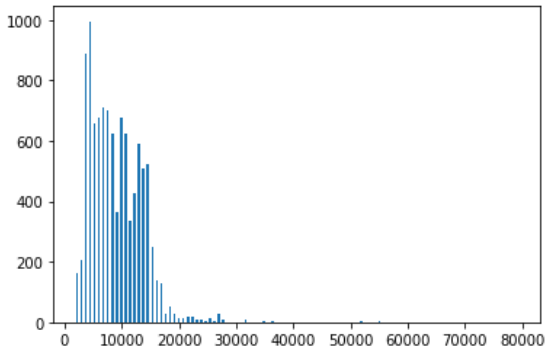
10225 rows × 30 columns

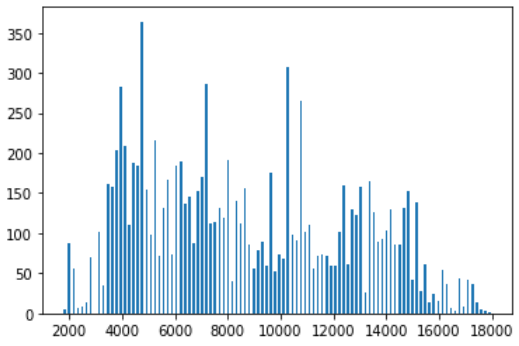
Затем была проведена нормализация:

	Date_of_Journey	Route	Dep_Time	Arrival_Time	Duration	Total_Stops	Source_Chennai	Source_Delhi	Source_Kolkata	Source_Mumbai	...	Multiple carriers
0	0.198276	0.151261	0.932862	0.045455	0.038153	0.00	0.0	0.0	0.0	0.0	...	0.0
1	0.517241	0.655462	0.233216	0.552448	0.148594	0.50	0.0	0.0	1.0	0.0	...	0.0
2	0.844828	0.924370	0.385159	0.181818	0.427711	0.50	0.0	1.0	0.0	0.0	...	0.0
3	0.612069	0.714286	0.752650	0.982517	0.100402	0.25	0.0	0.0	1.0	0.0	...	0.0
4	0.000000	0.243697	0.699647	0.902098	0.084337	0.25	0.0	0.0	0.0	0.0	...	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...
10678	0.327586	0.487395	0.830389	0.937063	0.030120	0.00	0.0	0.0	1.0	0.0	...	0.0
10679	0.482759	0.487395	0.865724	0.975524	0.032129	0.00	0.0	0.0	1.0	0.0	...	0.0
10680	0.482759	0.151261	0.339223	0.472028	0.042169	0.00	0.0	0.0	0.0	0.0	...	0.0
10681	0.000000	0.151261	0.473498	0.590909	0.034137	0.00	0.0	0.0	0.0	0.0	...	0.0
10682	0.586207	0.848739	0.448763	0.804196	0.170683	0.50	0.0	1.0	0.0	0.0	...	0.0

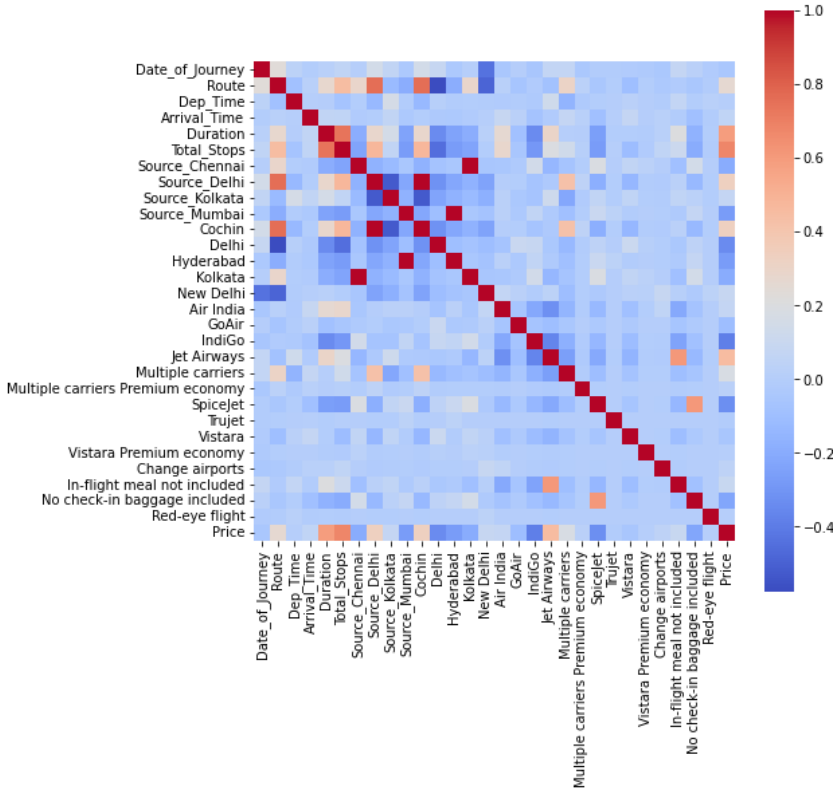
10225 rows × 29 columns

Распределение цены полета до и после обработки:





Корреляции:



4. Обучение линейной регрессии. Результаты.

Данные были поделены на обучающую и тестовую выборки в соотношении 70:30:

```
y_train.shape, y_test.shape
```

```
((7157,), (3068,))
```

```
X_train.shape, X_test.shape
```

```
((7157, 29), (3068, 29))
```

Обучение:

```
model = LinearRegression().fit(X_train, y_train)
```

Для оценки результатов были использованы 3 метрики: R2, MAE и MSE:

$$R2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})}, \qquad \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \qquad MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Были получены следующие значения метрик:

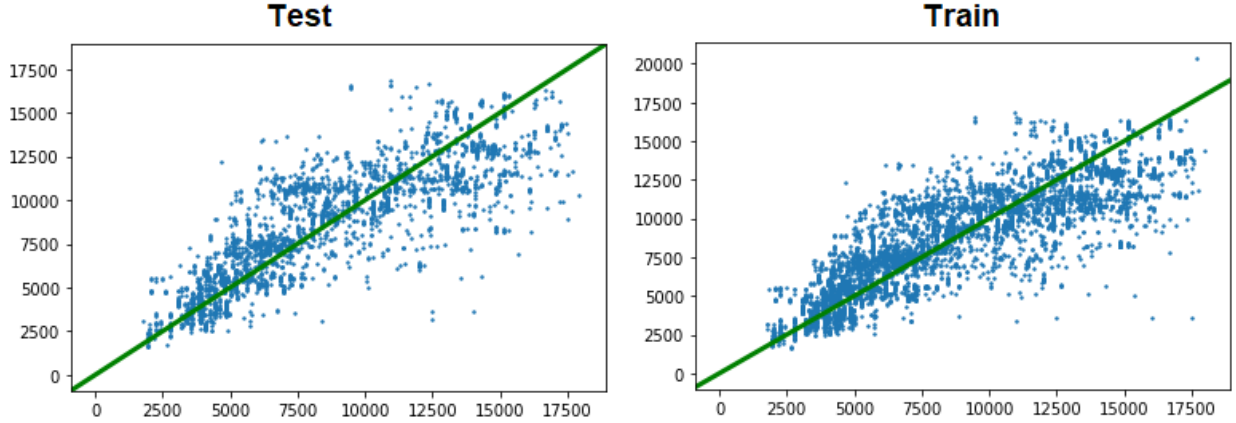
```
R2 train:   0.7215851937277724
R2 test:    0.7082920323244801

MAE train:  1553.8804931640625
MAE test:   1555.5675048828125

MSE train   4203709.5
MSE test:   4323156.5
```

Судя по R2, модель слегка переучилась (но не критично). Результат R2=0.72 говорит о “сносной” работе модели (max R2 = 1).

Были построены графики, показывающие, насколько хорошо работает регрессия. По оси x отмечены настоящие цены, по оси y - предсказанные. В идеале должна получиться зеленая прямая, но в реальности, конечно, такого не бывает.



5. Выводы

В результате работы была решена задача линейной регрессии для предсказания цены полета по 10 признакам. Для этого была проделана большая работа по обработке исходных данных. Обработка позволила удвоить метрику R2. Несмотря на то, что линейная регрессия - один из самых примитивных методов машинного обучения, удалось получить нормальные результаты - R2=0.70.