Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Московский государственный технический университет
имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ _____ИНФОРМАТИКА И СИСТЕМЫ УПРАВЛЕНИЯ_____

КАФЕДРА _____СИСТЕМЫ ОБРАБОТКИ ИНФОРМАЦИИ И УПРАВЛЕНИЯ_____

РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

К КУРСОВОЙ РАБОТЕ

НА ТЕМУ:

Разведочный анализ данных для выбранного датасета

Студент __ИУ5-14М_____ _____ _____Н.В. Журавлев_____
(Группа) (Подпись, дата) (И.О.Фамилия)

Руководитель курсовой работы _____ _____В.Ю. Строганов____
(Подпись, дата) (И.О.Фамилия)

2023 г.

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

УТВЕРЖДАЮ
Заведующий кафедрой ___ИУ5____
(Индекс)
_____ ___В.И. Терехов__
(И.О.Фамилия)
«_04_» ____сентября_____ 2023 г.

ЗАДАНИЕ
на выполнение курсовой работы

по дисциплине ____Многомерный анализ данных в системах искусственного интеллекта _____
_____

Студент группы _ИУ5-14М__

Журавлев Николай Вадимович_____
(Фамилия, имя, отчество)

Тема курсовой работы Разведочный анализ данных для выбранного датасета_____
_____

Направленность КР (учебная, исследовательская, практическая, производственная, др.)
_____УЧЕБНАЯ_____
Источник тематики (кафедра, предприятие, НИР) _____КАФЕДРА_____

График выполнения работы: 25% к ___ нед., 50% к ___ нед., 75% к __ нед., 100% к ___ нед.

Задание _Произвести разведочный анализ данных выбранного датасета_____
_____

Оформление курсовой работы:

Расчетно-пояснительная записка на __11_ листах формата А4.

_____
_____
_____
_____

Дата выдачи задания «_04_» ____сентября_____ 2023 г.

Руководитель курсовой работы _____ ____В.Ю. Строганов_____
(Подпись, дата) (И.О.Фамилия)

Студент _____ _____Н.В.Журавлев__
(Подпись, дата) (И.О.Фамилия)

Примечание: Задание оформляется в двух экземплярах: один выдается студенту, второй хранится на кафедре.

Оглавление

Цель работы

Целью данной курсовой работы является апробация методов разведочного анализа данных и прогнозирования:

- корреляционный анализ
- дисперсионный анализ
- регрессионный анализ
- кластерный анализ
- дискриминантный анализ

Описание анализируемого датасета

В качестве исследуемого набора выбран Estate_valuation.csv – набор рыночных данных по оценке недвижимости, собранный в Синдианском округе, город Нью-Тайбэй, Тайвань. Набор содержит:

1. X1 transaction date – дата заключения сделки; значения - дата в формате год, прошедшая часть (например, 2013,250 = март 2013 г., 2013,500 = июнь 2013)
2. X2 house age – возраст дома; значения – любые численные
3. X3 distance to the nearest MRT – метров до ближайшей станции метро; значения - любые численные
4. X4 number of convenience stores - количество магазинов повседневного спроса в шаговой доступности; значения - любые численные
5. X5 latitude – широта, на которой расположен дом; значения - любые численные
6. X6 longitude – долгота, на которой расположен дом; значения - любые численные
7. Y house price of unit area - стоимость дома за единицу площади; значения - любые численные

Корреляционный анализ

Корреляционный анализ - это статистический инструмент, позволяющий установить связь между двумя или более различными переменными, а также оценить, насколько сильна взаимосвязь между этими переменными. Суть корреляционного анализа заключается в вычислении коэффициента корреляции, который показывает степень линейной зависимости между двумя переменными.

Если результат корреляционного анализа положительный, то взаимосвязь двух переменных прямо пропорциональная. Это означает, что при увеличении одной переменной, вторая будет также увеличиваться. Как правило, такой результат принято называть “позитивной корреляцией”. Если результат корреляционного анализа отрицательный, то взаимосвязь двух переменных обратно пропорциональная. Это означает, что при увеличении одной переменной, вторая будет уменьшаться. Такой эффект называется “отрицательной корреляцией”. Если результат анализа стремится к нулю, то взаимосвязь между двумя переменными отсутствует.

В данной курсовой работе использовался корреляционный анализ Спирмена.

Листинг 1

```
df = pd.read_csv('csvwn.csv', encoding='latin-1', sep=';')
numeric = ['X1', 'X2', 'X3', 'X4', 'X5', 'X6', 'Y']
sns.pairplot(df[numeric])
print(df[numeric].corr(method='spearman', numeric_only=False))
```

В результате получаем значения показанные на рис.1.

	X1	X2	X3	X4	X5	X6	Y
X1	1.000000	0.034686	0.093315	0.002647	0.023612	-0.013735	0.066928
X2	0.034686	1.000000	0.129605	0.009240	0.042142	-0.109674	-0.281753
X3	0.093315	0.129605	1.000000	-0.688165	-0.425107	-0.469393	-0.775948
X4	0.002647	0.009240	-0.688165	1.000000	0.429476	0.412054	0.617333
X5	0.023612	0.042142	-0.425107	0.429476	1.000000	0.264922	0.578419
X6	-0.013735	-0.109674	-0.469393	0.412054	0.264922	1.000000	0.437672
Y	0.066928	-0.281753	-0.775948	0.617333	0.578419	0.437672	1.000000

Рисунок 1

Интерпретация значений коэффициента корреляции:

1. до 0,5 – слабая корреляции
2. от 0,5 до 0,7 – средняя корреляции
3. от 0,7 до 0,9 – высокая корреляции
4. от 0,9 - очень высокая корреляция

Исходя из полученных результатов можно сделать следующие выводы:

1. Стоимость дома за единицу площади в значительной степени зависит от количества магазинов повседневного спроса в шаговой доступности (X4 number of convenience stores)
2. Стоимость дома за единицу площади так же сильно зависит от метров до ближайшей станции метро (X3 distance to the nearest MRT)
3. В меньшей, но значительной степени стоимость дома за единицу площади зависит от широты и долготы, на которых расположен дом (X5 latitude, X6 longitude)

Остальные зависимости не рассматриваются из-за за того, что рассматривается столбец Y house price of unit area.

Регрессионного анализа — линейная регрессия

Для регрессионного анализа выдвигаем гипотезу о том, какие факторы влияют на стоимость дома за единицу площади. Гипотеза для регрессионного анализа сформулирована следующим образом: "Стоимость дома за единицу

площади зависит от возраста дома, метров до ближайшей станции метро, количество магазинов повседневного спроса в шаговой доступности, широты, на которой расположен дом, долготы, на которой расположен дом".

Листинг 2

```
X = df[['X1', 'X2', 'X3', 'X4', 'X5', 'X6']]

Y = df['Y']

X = sm.add_constant(X, prepend=False)

model = OLS(Y, X)

res = model.fit()

print(res.summary())
```

После выполнения получается результат представленный на рис.2.

Dep. Variable:	Y	R-squared:	0.582
Model:	OLS	Adj. R-squared:	0.576
Method:	Least Squares	F-statistic:	94.59
Date:	Fri, 12 Jan 2024	Prob (F-statistic):	4.86e-74
Time:	11:55:06	Log-Likelihood:	-1487.0
No. Observations:	414	AIC:	2988.
Df Residuals:	407	BIC:	3016.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
X1	5.1462	1.557	3.305	0.001	2.085	8.207
X2	-0.2697	0.039	-7.000	0.000	-0.345	-0.194
X3	-0.0045	0.001	-6.250	0.000	-0.006	-0.003
X4	1.1333	0.188	6.023	0.000	0.763	1.503
X5	225.4730	44.567	5.059	0.000	137.863	313.083
X6	-12.4236	48.582	-0.256	0.798	-107.927	83.079

Рисунок 2

Значение R-squared всегда находится в диапазоне от 0 до 1, где 0 означает отсутствие связи между переменными, а 1 означает полную линейную связь. В нашем случае R-squared: 0.582, что означает среднюю линейную связь.

Столбец X6 может отбросить, так как коэффициент значимости p_value выше 0.05.

Dep. Variable:	Y	R-squared:	0.582
Model:	OLS	Adj. R-squared:	0.577
Method:	Least Squares	F-statistic:	113.8
Date:	Fri, 12 Jan 2024	Prob (F-statistic):	4.47e-75
Time:	12:05:14	Log-Likelihood:	-1487.0
No. Observations:	414	AIC:	2986.
Df Residuals:	408	BIC:	3010.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
X1	5.1348	1.555	3.303	0.001	2.079	8.191
X2	-0.2694	0.038	-7.003	0.000	-0.345	-0.194
X3	-0.0044	0.000	-8.887	0.000	-0.005	-0.003
X4	1.1361	0.188	6.056	0.000	0.767	1.505
X5	226.8816	44.174	5.136	0.000	140.044	313.719

Рисунок 3

После удаления X6 коэффициент детерминации не изменился. Это означает, что X6 не влияет на изменение зависимой переменной.

Дисперсионный анализ

В проведенном дисперсионном анализе используется гипотеза о равенстве средних значений в нескольких группах. В данном случае мы предполагаем, что средние значения измеряемых переменных (издержки, количество изделий, стоимость сборки и т.д.) одинаковы для всех групп или обработок. Формально, нулевая гипотеза (H0) в дисперсионном анализе звучит как "Средние значения всех групп равны".

Листинг 3

```
data = np.array(df[numeric])

f_value, p_value = stats.f_oneway(*data)

print("F-значение:", f_value)

print("p-значение:", p_value)
```

После проведения дисперсионного анализа, и получения значений F-тест и p-значения, мы можем принять решение об отклонении или не отклонении нулевой гипотезы в зависимости от уровня значимости (обычно 0.05).

Для проведения дисперсионного анализа используется функцию f_oneway

из библиотеки scipy.stats, которая возвращает значения F-тест и p-значения.

F-значение: 0.26574880853703486

p-значение: 0.9999999999999999

Так как p-значение больше 0.05, то нулевую гипотезу не отвергаем.

Кластерный анализ

Для проведения кластерного анализа, используем метод k-средних из библиотеки scikit-learn для Python.

Листинг 4

```
df1 = pd.DataFrame(df[numeric])
scaler = StandardScaler()
scaled_data = scaler.fit_transform(df1)
kmeans = KMeans(n_clusters=6)
kmeans.fit(scaled_data)
df1['Cluster'] = kmeans.labels_
print(df1)
```

В результате получаем распределение по кластерам:

	X1	X2	X3	X4	X5	X6	Y	Cluster
0	2012.917	32.0	84.87882	10	24.98298	121.54024	37.9	4
1	2012.917	19.5	306.59470	9	24.98034	121.53951	42.2	0
2	2013.583	13.3	561.98450	5	24.98746	121.54391	47.3	2
3	2013.500	13.3	561.98450	5	24.98746	121.54391	54.8	2
4	2012.833	5.0	390.56840	5	24.97937	121.54245	43.1	0
..	...	...	...	..	...	...	...	...
409	2013.000	13.7	4082.01500	0	24.94155	121.50381	15.4	3
410	2012.667	5.6	90.45606	9	24.97433	121.54310	50.0	0
411	2013.250	18.8	390.96960	7	24.97923	121.53986	40.6	2
412	2013.000	8.1	104.81010	5	24.96674	121.54067	52.5	0
413	2013.500	6.5	90.45606	9	24.97433	121.54310	63.9	2

Дискриминантный анализ

Дискриминантный анализ (Discriminant Analysis) - это статистический метод, для решения задач распознавания образов, который используется для принятия решения о том, какие переменные разделяют возникающие наборы

данных (так называемые «группы»). В отличие от кластерного анализа в дискриминантном анализе группы известны заранее. Основная цель дискриминантного анализа состоит в том, чтобы определить, какие переменные наилучшим образом разделяют эти группы.

Суть дискриминантного анализа заключается в поиске линейной комбинации признаков, которая наилучшим образом разделяет группы объектов, которая называется дискриминантной функцией. После того как она найдена, её используют для предсказания принадлежности новых объектов к определенным группам.

Листинг 5

```
data = {
   'transaction_date': df['X1'],
   'house_price_of_unit_area': df['Y'],
   'house_age': df['X2'],
   'distance_to_the_nearest_MRT': df['X3'],
   'number_of_convenience_stores': df['X4'],
   'latitude': df['X5'],
   'longitude': df['X6']
}
df2 = pd.DataFrame(data)
X = df2[['house_age', 'distance_to_the_nearest_MRT', 'number_of_convenience_stores']]
y = df2[['house_price_of_unit_area']]
y = y.astype('int')
model = LinearDiscriminantAnalysis()
model.fit(X, y)
new_data = [[19.5, 306.5947, 9]]
predicted_class = model.predict(new_data)
print(f'Предсказание: {predicted_class[0]}')
```

Для проверки предсказания нового значения использовалось вторая строка из датасета. Из дискриминантного анализа, мы получаем предсказание для цены

равным 44. При сравнении с изначальным датасетом, где результатом является 42.2, можно увидеть, что это является близко к полученному результату.

Заключение

В ходе выполнения данной курсовой работы были проанализированы данные датасета рыночных данных по оценке недвижимости. Из результатов корреляционного анализа можно сделать следующие выводы:

1. Стоимость дома за единицу площади в значительной степени зависит от количества магазинов повседневного спроса в шаговой доступности (X4 number of convenience stores)
2. Стоимость дома за единицу площади так же сильно зависит от метров до ближайшей станции метро (X3 distance to the nearest MRT)
3. В меньшей, но значительной степени стоимость дома за единицу площади зависит от широты и долготы, на которых расположен дом (X5 latitude, X6 longitude)

Результат регрессионного анализа – нахождения значения R-squared равным 0.582, что означает среднюю линейную связь.

Результаты кластерного анализа показали разделение данных на 6 кластеров.

Результаты дисперсионного анализа показали, что мы не отвергаем нулевую гипотезу: нет оснований считать средние значения различными.

Из дискриминантного анализа, мы получаем предсказание цены для второй строки датасета равным 44. При сравнении с изначальным датасетом, где результатом является 42.2, можно увидеть, что это является близко к полученному результату.

Список литературы

1. Афонин П.Н., Афонин Д.Н. Статистический анализ с применением современных программных средств : учебное пособие / Афонин П.Н., Афонин Д.Н; – СПб.: ИЦ «Интермедия», 2017. – 100 с.
2. Боровиков В.П. Statistica: искусство анализа данных на компьютере. Для профессионалов. СПб.: Питер, 2003. – 688 с.
3. Боровиков В.П., Ивченко Г.И. Прогнозирование в системе Statistica в среде Windows. Основы теории и интенсивная практика на компьютере. М., 2006. – 368 с
4. Вуколов Э.А. Основы статистического анализа. Практикум по статистическим методам и исследованию операций с использованием пакетов Statistica и Excel. М.: Форум, 2010. – 464 с.
5. Глинский В.В., Ионин В.Г. Статистический анализ. Учебное пособие. Новосибирск: Инфра-М, 2002. – 240 с
6. Симчера В.М. Методы многомерного анализа статистических данных. М.: Финансы и статистика. – 2008. – 400 с.
7. Тихомиров Н.П., Тихомирова Т.М., Ушмаев О.С. Методы эконометрики и многомерного статистического анализа. М.: Экономика, 2011. – 640 с.
8. Халафян А.А. Statistica 6. Статистический анализ данных. СПб.: Бином-Пресс, 2010. – 528 с.