Лабораторная работа №5

Выполнил: Величко Максим Иванович, М32061

Весь код для удобства выложен на мой GitHub(<https://github.com/maksve11/ITMO_MatStat_sem4/tree/main/lab5>)

Изображение выглядит как текст, Шрифт, линия, снимок экрана

Автоматически созданное описание

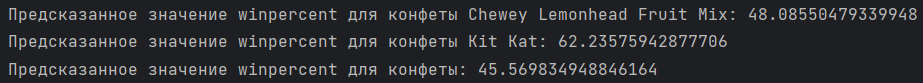
import pandas as pd  
from sklearn.linear\_model import LinearRegression  
import statsmodels.api as sm  
  
# Загрузка данных из файла  
data = pd.read\_csv('candy-data.csv', delimiter=',')  
  
# Удаление конфет Chewey Lemonhead Fruit Mix и Kit Kat  
data = data[(data['competitorname'] != 'Chewey Lemonhead Fruit Mix') & (data['competitorname'] != 'Kit Kat')]  
  
# Определение предикторов и отклика  
predictors = ['chocolate', 'fruity', 'caramel', 'peanutyalmondy', 'nougat', 'crispedricewafer', 'hard', 'bar', 'pluribus', 'sugarpercent', 'pricepercent']  
response = 'winpercent'  
  
# Разделение данных на предикторы и отклик  
X = data[predictors]  
y = data[response]  
  
# Создание и обучение модели линейной многомерной регрессии  
model = LinearRegression()  
model.fit(X, y)

В этом участке кода мы обучили нашу модель линейной регрессии, разделив нашу выборку данных на предикаторы и отклики.

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

# Предсказание значения winpercent для конфеты Chewey Lemonhead Fruit Mix  
chewey\_lemonhead = [[0, 1, 0, 0, 0, 0, 0, 0, 1, 0.73199999, 0.51099998]]  
predicted\_winpercent = model.predict(chewey\_lemonhead)  
  
print("Предсказанное значение winpercent для конфеты Chewey Lemonhead Fruit Mix:", predicted\_winpercent[0])  
  
kit\_kat = [[1, 0, 0, 0, 0, 1, 0, 1, 0, 0.31299999, 0.51099998]]  
predicted\_winpercent = model.predict(kit\_kat)  
  
print("Предсказанное значение winpercent для конфеты Kit Kat:", predicted\_winpercent[0])  
  
candy = [[0, 0, 0, 1, 0, 1, 0, 1, 1, 0.543, 0.402]]  
predicted\_winpercent = model.predict(candy)  
print("Предсказанное значение winpercent для конфеты:", predicted\_winpercent[0])

Т. к. по условию задания, нам нельзя учитывать конфеты Chewey Lemonhead Fruit Mix и Kit Kat, то данные для предсказания значения winpercent, мы берем из нашего файла candy-data.csv.  
И получаем результаты:  


Т. к. наши ответы совпали с платформой openedu, значит мы обучили нашу модель правильно.

Нужно построить доверительные интервалы для наших предикторов.

Для построения доверительных интервалов для коэффициентов модели линейной регрессии и оценки статистической значимости этих коэффициентов можно использовать библиотеку **statsmodels** помимо **sklearn.linear\_model**.

X = sm.add\_constant(X)  
model = sm.OLS(y, X)  
results = model.fit()  
  
print(results.summary())

Чтобы оценить свободный член нашей модели, мы добавляем константу к предикторам, в итоге получаем такой результат:  
Изображение выглядит как текст, снимок экрана

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, меню

Автоматически созданное описание

Результаты сводной статистики модели линейной регрессии (OLS Regression Results) позволяют оценить значимость и влияние каждой переменной на отклик (**winpercent**). Рассмотрим основные показатели:

* R-squared (R-квадрат): 0.587 R-квадрат представляет собой меру объясненной дисперсии моделью. Значение 0.587 означает, что модель объясняет около 58.7% вариации в отклике.
* Adj. R-squared (скорректированный R-квадрат): 0.506 Скорректированный R-квадрат учитывает количество предикторов и позволяет сравнивать модели с разным числом предикторов. Здесь значение составляет 50.6%.
* F-statistic (F-статистика): 7.234 F-статистика используется для проверки гипотезы о значимости модели в целом. Здесь F-статистика равна 7.234, а соответствующее p-значение составляет 1.70e-07 (очень маленькое значение). Это говорит о статистической значимости модели в целом.
* Доверительные интервалы (0.025, 0.975): Доверительные интервалы показывают диапазон значений, в котором с определенной вероятностью находится истинное значение коэффициента.

Нужно проверить некоторые гипотезы:  
Есть подозрение, что чем больше sugarpercent, тем больше winpercent;

Есть подозрение, что чем hard меньше, тем больше winpercent;

Есть подозрение, что от количества chocolate зависит winpercent;  
Есть подозрение, что одновременно chocolate, caramel, nougat влияют на winpercent;

Для проверки гипотез о влиянии переменных **sugarpercent**, **hard** и **chocolate** на **winpercent** в модели линейной регрессии, можно использовать статистические тесты и интерпретацию значимости соответствующих коэффициентов модели.

В **statsmodels** результаты модели **results** уже содержат информацию о значимости коэффициентов, выведенную в **results.summary()**. Давайте рассмотрим эти тесты для проверки гипотез:

1. Гипотеза: чем больше **sugarpercent**, тем больше **winpercent**.
   * Значимость коэффициента **sugarpercent** будет показывать, есть ли статистически значимая связь между **sugarpercent** и **winpercent**. Если p-значение (P>|t|) для **sugarpercent** меньше выбранного уровня значимости (обычно 0.05), то можно сделать вывод, что есть статистически значимая связь между **sugarpercent** и **winpercent**.
2. Гипотеза: чем **hard** меньше, тем больше **winpercent**.
   * Значимость коэффициента **hard** будет указывать, есть ли статистически значимая связь между **hard** и **winpercent**. Если p-значение (P>|t|) для **hard** меньше выбранного уровня значимости, то можно сделать вывод, что есть статистически значимая связь между **hard** и **winpercent**.
3. Гипотеза: Зависимость **winpercent** от количества **chocolate**.
   * Значимость коэффициента **chocolate** указывает, есть ли статистически значимая связь между наличием **chocolate** и **winpercent**. Если p-значение (P>|t|) для **chocolate** меньше выбранного уровня значимости, то можно сделать вывод, что есть статистически значимая связь между наличием **chocolate** и **winpercent**.

Дополнительно можно рассмотреть интервалы доверия для соответствующих коэффициентов, чтобы оценить их значимость и диапазон возможных значений.

Для проверки последней гипотезы о влиянии одновременно переменных **chocolate**, **caramel** и **nougat** на **winpercent** можно воспользоваться статистическим тестом на значимость коэффициентов данных переменных в множественной регрессии.

В **statsmodels** можно использовать тест F-статистики (F-test) для проверки значимости всей группы переменных одновременно. Гипотеза заключается в том, что все коэффициенты этих переменных равны нулю (нет влияния).

Вот код, решающий наши задачи:

# Проверка гипотез  
  
# Гипотеза: Чем больше sugarpercent, тем больше winpercent  
sugar\_coef = results.params['sugarpercent']  
sugar\_p\_value = results.pvalues['sugarpercent']  
print("Значимость коэффициента sugarpercent: p =", sugar\_p\_value)  
if sugar\_p\_value < 0.05:  
 print("Гипотеза о влиянии sugarpercent на winpercent подтверждена.")  
else:  
 print("Нет достаточных доказательств в пользу гипотезы о влиянии sugarpercent на winpercent.")  
  
# Гипотеза: Чем hard меньше, тем больше winpercent  
hard\_coef = results.params['hard']  
hard\_p\_value = results.pvalues['hard']  
print("Значимость коэффициента hard: p =", hard\_p\_value)  
if hard\_p\_value < 0.05:  
 print("Гипотеза о влиянии hard на winpercent подтверждена.")  
else:  
 print("Нет достаточных доказательств в пользу гипотезы о влиянии hard на winpercent.")  
  
# Гипотеза: Зависимость winpercent от количества chocolate  
chocolate\_coef = results.params['chocolate']  
chocolate\_p\_value = results.pvalues['chocolate']  
print("Значимость коэффициента chocolate: p =", chocolate\_p\_value)  
if chocolate\_p\_value < 0.05:  
 print("Гипотеза о влиянии chocolate на winpercent подтверждена.")  
else:  
 print("Нет достаточных доказательств в пользу гипотезы о влиянии chocolate на winpercent.")  
  
# Гипотеза: Одновременное влияние chocolate, caramel, nougat на winpercent  
  
hypotheses = '(chocolate = 0), (caramel = 0), (nougat = 0)'  
f\_test = results.f\_test(hypotheses)  
  
print("Значимость теста F-статистики:", f\_test.pvalue)  
if f\_test.pvalue < 0.05:  
 print("Гипотеза о влиянии chocolate, caramel, nougat на winpercent подтверждена.")  
else:  
 print("Нет достаточных доказательств в пользу гипотезы о влиянии chocolate, caramel, nougat на winpercent.")

В результате выполнения кода будет выведена сводная статистика модели, а затем значения значимости и интерпретации для каждого из предполагаемых влияющих факторов **sugarpercent**, **hard** и **chocolate**, также будет выведена сводная статистика модели, а затем результаты теста F-статистики для проверки гипотезы о влиянии одновременно переменных **chocolate**, **caramel** и **nougat** на **winpercent**. Если p-значение (p-value) теста F-статистики меньше выбранного уровня значимости (обычно 0.05), то можно сделать вывод, что есть статистически значимое одновременное влияние данных переменных на **winpercent**.

Вот результаты:  
Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание

Исходя из результатов, можно сделать следующие выводы:

1. Значимость модели:
   * F-статистика говорит о статистической значимости модели в целом, что подтверждает наличие связи между предикторами и откликом.
2. Значимость предикторов:
   * Переменная **chocolate** имеет сильное и положительное влияние на **winpercent**, со статистически значимым коэффициентом.
   * Переменная **sugarpercent** также имеет положительное влияние на **winpercent** и статистическую значимость.
   * Переменная **hard** имеет отрицательное влияние на **winpercent** и статистическую значимость.
3. Незначимость предикторов:
   * Переменные **caramel**, **nougat**, **crispedricewafer**, **bar** и **pluribus** не имеют статистически значимого влияния на **winpercent**.

На основе этих результатов можно сделать выводы:

* В данных предоставленной модели наиболее значимыми предикторами **winpercent** являются **chocolate** и **sugarpercent**. Увеличение значения этих переменных обычно сопровождается повышением **winpercent**.
* Переменная **hard** имеет негативное влияние на **winpercent**. То есть, снижение значения **hard** может привести к увеличению **winpercent**.
* Влияния одновременно переменных **caramel**, **nougat** и **chocolate** доказано на значение **winpercent**.