МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное

образовательное учреждение высшего образования

«ЧЕРЕПОВЕЦКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

Институт информационных технологий

*наименование института(факультета)*

Кафедра математического и программного обеспечения ЭВМ

*наименование кафедры*

Основы Data Science

*наименование дисциплины в соответствии с учебным планом*

ЛАБОРАТОРНАЯ РАБОТА №4

«Изучение связи между переменными»

|  |  |
| --- | --- |
| Исполнитель |  |
| Студент | 1ПИб-02-3оп-22 |
|  | *группа* |
|  | Маркелов С. А. |
|  | *ФИО* |
| Руководитель | Юдина О.В. |
|  | *ФИО преподавателя* |
| Оценка |  |
| Подпись |  |

2024 год

Цель: Знакомство с этапом понимания данных стандарта CRISP-DM.

Задача этапа – исследовать связи между переменными между переменными.

Задание

Для обнаружения зависимостей воспользуйтесь графическими иллюстрациями, а также результатами предыдущих работ

1. Выделите переменные, для которых может быть посчитана параметрическая корреляция (Пирсона). Посчитайте. Оцените значимость коэффициента Пирсона.
2. Выберите зависимости, для которых можно воспользоваться ранговым коэффициентом Спирмена. Вычислите те, которые могут быть использованы при описании данных. Вычислите и оцените коэффициент.
3. Подберите подходящие инструменты для нелинейных зависимостей. Если такие есть в наборе данных – вычислите коэффициенты, оцените результаты.
4. Оцените статистическую значимость найденных зависимостей.

Воспользовавшись программой Excel, постройте регрессионную зависимость. Для построения зависимости может быть использована часть выборки, группы записей, для которых вы установили закономерности. Попробуйте использовать качественную переменную.

Для анализа результатов воспользуйтесь оценками Excel. Поясните оценки и результаты. Сделайте выводы о пригодности регрессии, обоснуйте их.

1. Корреляция Пирсона

Корреляция Пирсона применяется для непрерывных количественных переменных. Таких переменных в наборе данных 2 – возраст и количество часов сна

Рассчитаем корреляцию с помощью функции =PEARSON: -0,0548

При значении коэффициента |0,0| + |±0,3| связь практически отсутствует

При значении коэффициента |0,3| + |±0,5| связь слабая

При значении коэффициента |0,5| + |±0,7| связь умеренная

При значении коэффициента |0,7| + |±1,0| связь сильная

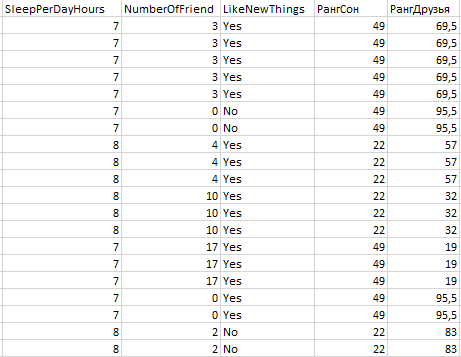
Следовательно, связь практически отсутствует

Проверим статистическую значимость:

|-0,54| < |1,66|, следовательно связь не является статистически значимой

1. Ранговый коэффициент Спирмена

Для нахождения коэффициента необходимо найти ранги переменных.



Находим корреляцию с помощью функции =КОРРЕЛ: 0,1334. Следовательно, связь практически отсутствует

Проверим статистическую значимость:

|1,33| < |1,66|, следовательно связь не является статистически значимой

Контрольные вопросы:

1. Когда может быть использован коэффициент корреляции Пирсона?

Коэффициент корреляции Пирсон может быть использован для непрерывных линейных зависимостей, имеющих нормальное распределение и не имеющих выбросов

1. Что такое ранговая корреляция? Когда она применяется?

Ранговая корреляция – это статистическая мера, которая оценивает степень силы и направления связи между двумя переменными, основываясь на рангах значений этих переменных, а не на их фактических значениях. В отличие от коэффициента корреляции Пирсона, который измеряет линейную связь между переменными, ранговые корреляции оценивают силу и направление монотонной связи.

Существует несколько различных коэффициентов ранговой корреляции, но два наиболее распространенных – это коэффициенты ранговой корреляции Спирмена и Кендалла.

Ранговая корреляция применяется в тех случаях, когда данные не соответствуют требованиям для использования коэффициента корреляции Пирсона, например, когда они не имеют нормального распределения или если имеются выбросы. Она также полезна при работе с порядковыми или ранговыми данными, когда исследуются связи между ранжированными переменными.

1. Когда применяется коэффициент Фехнера? Как им воспользоваться?

Коэффициент Фехнера применяется для измерения тесноты линейной связи между двумя переменными.

Коэффициент Фехнера рассчитывается по формуле , где С – число совпадений, Н – число несовпадений знаков.

Коэффициент Фехнера может принимать значения от -1 до 1. Значения близкие к 1 указывают на сильную связь между переменными, значения близкие к 0 указывают на отсутствие связи, а отрицательные значения указывают на противоположную связь (то есть на обратную корреляцию).

1. На основания чего может проводится анализ регрессии?

Анализ регрессии проводится на основании данных, которые представляют собой наблюдения для двух или более переменных. Основание для анализа регрессии включает следующие компоненты:

* Зависимая переменная – это переменная, которая должна быть предсказана или объяснена. Например, если мы исследуем факторы, влияющие на продажи автомобилей, количество проданных автомобилей может быть зависимой переменной.
* Независимые переменные – это переменные, которые используются для предсказания или объяснения изменений в зависимой переменной. Например, в анализе продаж автомобилей независимыми переменными могут быть цена автомобиля, рекламные расходы, климатические условия и т. д.

1. Какое значение имеют остатки?

Остатки – это разница между фактическими значениями зависимой переменной и значениями, предсказанными моделью регрессии для соответствующих наблюдений.

Остатки имеют следующее значение:

* Проверка адекватности модели: Остатки используются для оценки того, насколько хорошо модель соответствует данным. Если модель хорошо подходит для данных, остатки должны быть случайно распределены вокруг нуля. Если остатки демонстрируют систематические отклонения от нуля, это может указывать на недостаточность модели.
* Оценка точности предсказаний: Остатки также могут быть использованы для оценки точности предсказаний модели. Меньшие остатки обычно указывают на более точные предсказания, тогда как большие остатки могут указывать на недооценку или переоценку модели.
* Выявление нарушений предположений: Анализ остатков может помочь выявить нарушения предположений о модели, такие как наличие нелинейности, выбросов или влиятельных наблюдений.

Вывод:

В ходе выполнения лабораторной работы были исследованы взаимосвязи переменных с помощью коэффициентов Пирсона и Спирмена.

В результате расчетов выяснилось, что коэффициенты для переменных «Возраст» и «Количество часов сна» и переменных «Количество часов сна» и «Количество друзей» не являются статистически значимыми.