|  |  |
| --- | --- |
|  | **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 Теоретическая информатика и компьютерные технологии

**Отчет по лабораторной работе № 4**

***по курсу «Моделирование»***

**«Алгоритм обработки и анализа данных при моделировании прибыльности»**

Студент ИУ9-81Б **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**  Костриця М.И.

(Группа) (Подпись, дата) (Фамилия И.О.)

Преподаватель **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** Домрачева А. Б. (Подпись, дата) (Фамилия И.О.)

*Москва, 2023 г.*

# Цель работы

Предоставлены пробы нефти в трёх регионах: в каждом 100 000 месторождений, где измерили качество нефти и объём её запасов. Необходимо построить модель, которая поможет определить регион, где добыча принесёт наибольшую прибыль.

Шаги для выбора локации:

1. в избранном регионе ищут месторождения, для каждого определяют значения признаков;
2. строят модель и оценивают объём запасов;
3. выбирают месторождения с самым высокими оценками значений, количество месторождений зависит от бюджета компании и стоимости разработки одной скважины;
4. прибыль равна суммарной прибыли отобранных месторождений.

Предоставлены три датасета, соответствующие трем разным исследуемым локациям, в них

1. id — уникальный идентификатор скважины;
2. f0, f1, f2 — три признака точек (неважно, что они означают, но сами признаки значимы);
3. product — объём запасов в скважине (тыс. баррелей).

Необходимо реализовать алгоритм поиска наиболее выгодного региона для разработки скважин и добычи нефти на основе входных данных, включающих список скважин с признаками и количеством сырья.

Для решения поставленной задачи использовалась модель линейной регрессии (остальные — недостаточно предсказуемые). При разведке региона исследуют 500 точек, из которых с помощью машинного обучения выбирают 16 лучших для разработки. Бюджет на разработку скважин в регионе — 500000 рублей. При нынешних ценах один баррель сырья приносит 450 рублей дохода. Доход с каждой единицы продукта составляет 450 тыс. рублей, поскольку объём указан в тысячах баррелей. После оценки рисков нужно оставить лишь те регионы, в которых вероятность убытков меньше 2.5%. Среди них выбирают регион с наибольшей средней прибылью.

# Теоретическая часть

Необходимо провести предварительную обработку данных. Выявить выбросы (если есть), рассчитать квартили, интерквартильный размах, выборочную дисперсию для всех столбцов каждого датасета. Определить корреляцию целевого признака (product) c зависимыми признаками для каждого дататсета.

*Медиана* — числовое значение признака, которое делит упорядоченную по возрастанию совокупность на две равных части.

*Квартили* — числовые значения признака, которые делят упорядоченную по возрастанию совокупность на четыре равных части. Раз квартили делят совокупность на четыре части, то квартилей бывает три варианта: первый (нижний), второй(средний), третий (верхний). Второй квартиль это и есть медиана.

*Интерквартильный размах* - интервал значений признака, содержащий центральные 50% наблюдений выборки, т.е. интервал между 25-м и 75-м процентилями.

*Выборочная дисперсия* – это среднее арифметическое значений вариантов части отобранных объектов генеральной совокупности (выборки).

*Генеральная совокупность* – это комплекс всех возможных объектов, относительно которых планируется вести наблюдение и формулировать выводы.

*Дисперсия выборки* или выборочная дисперсия оценивается по формуле:

*–* среднее значение выборки

*Корреляция* – статистическая взаимосвязь двух или более случайных величин, при этом изменения значений одной или нескольких из этих величин сопутствуют систематическому изменению значений другой или других величин.

*Линейная регрессия* — используемая в статистике регрессионная модель зависимости переменной 𝑦 от другой или нескольких других переменных (факторов, регрессоров, независимых переменных) 𝑥 с линейной функцией зависимости.

Цель линейной регрессии — поиск линии, которая наилучшим образом соответствует этим точкам. Функция потерь — это мера количества ошибок, которые линейная регрессия делает на наборе данных, которая вычисляет расстояние между предсказанным значением y(х) и его фактическим значением.

В данной лабораторной в качестве функции потерь использовалась среднеквадратичное отклонение (RMSE), для вычисления которого необходимо взять среднее от квадратов длин значения ошибок.

Также для более точного анализа данных использовался масштабатор (Scaler), который необходим для изменения размера распределения значений так, чтобы среднее значение наблюдаемых значений было равно 0, а стандартное отклонение – 1. Он необходим, когда характеристики входного набора данных сильно различаются между диапазонами.

Таким образом, масштабатор (Scaler) приводит данные к такому виду, чтобы все переменные в равной степени влияли на обучение модели, вне зависимости от диапазона значений каждой переменной.

Подготовка к расчёту прибыли:

1. Все ключевые значения для расчётов сохранить в отдельных переменных.
2. Рассчитать достаточный объём сырья для безубыточной разработки новой скважины. Сравнить полученный объём сырья со средним запасом в каждом регионе.

Написать функцию для расчёта прибыли по выбранным скважинам и предсказаниям модели:

1. Выбрать скважины с максимальными значениями предсказаний.
2. Просуммировать целевое значение объёма сырья, соответствующее этим предсказаниям.
3. Рассчитать прибыль для полученного объёма сырья.

Подсчёт рисков и прибыли для каждого региона:

* 1. Найти распределение прибыли. Выборки небольшие, поэтому есть смысл использовать технологию bootstrap.
  2. Найти среднюю прибыль, 95%-й доверительный интервал и риск убытков. Убыток — это отрицательная прибыль.
  3. Предложить регион для разработки скважин и обосновать выбор.

*Доверительный интервал* — термин, используемый в математической статистике при интервальной оценке статистических параметров, более предпочтительной при небольшом объёме выборки, чем точечная. Доверительным называют интервал, который покрывает неизвестный

параметр с заданной надёжностью.

Доверительным называется интервал, в который попадают измеренные в эксперименте значения, соответствующие доверительной вероятности.

1. **Практическая реализация**

Для анализа данных использовались библиотеки pandas и scikit-learn на языке программирования Python. Исследование было проведено на трех наборах данных, каждый из которых содержал информацию о 40 месторождениях. Данные были загружены из файлов с помощью библиотеки pandas.

Затем свойства f0, f1, f2 каждого месторождения были выделены в отдельные массивы для предсказания количества сырья в скважине и ожидаемого результата.

Для обучения модели использовалась линейная регрессия. Данные были разделены на обучающие и тестовые в соотношении 3:1. Модель была обучена на 75% данных, а на оставшихся

25% производился прогноз и сравнение с правильным значением для вычисления RMSE.

Для каждого региона был вычислен доверительный интервал для оценки рисков и прибыли с помощью технологии bootstrap - распределение статистик вероятностных распределений, основанный на многократной генерации выборок с использованием генератора случайных величин. Результаты показали, что первый регион является самым прибыльным с небольшим среднеквадратичным отклонением, второй регион - наименее прибыльный, а третий имеет высокое среднеквадратичное отклонение. Однако стоит отметить, что объем тестовых данных был небольшим, что могло повлиять на обучение модели и точность результатов

def remove\_ouliers(frame,column):

q25=np.array(frame[column].quantile(0.25))

q75=np.array(frame[column].quantile(0.75))

first\_part=q25-1.5\*(q75-q25)

second\_part=q75+1.5\*(q75-q25)

del\_index = []

for index\_value, value in zip(frame[column].index,frame[column]):

if second\_part <= value or value <= first\_part:

del\_index.append(index\_value)

print('Количество строк, выбранных для удаления: ',len(del\_index))

return del\_index

Функция для удаления данных, выходящих за границы

best\_model\_zero = data\_dict['pipelines']['zero\_frame']

best\_model\_first = data\_dict['pipelines']['first\_frame\_out\_f2']

best\_model\_second = data\_dict['pipelines']['second\_frame']

predicted\_values\_zero = best\_model\_zero.predict(data\_dict['valid']['zero\_frame'][0])

predicted\_values\_first = best\_model\_first.predict(data\_dict['valid']['first\_frame\_out\_f2'][0])

predicted\_values\_second = best\_model\_second.predict(data\_dict['valid']['second\_frame'][0])

RMSE\_model\_zero = (mean\_squared\_error(data\_dict['valid']['zero\_frame'][1],predicted\_values\_zero))\*\*0.5

RMSE\_model\_first = (mean\_squared\_error(data\_dict['valid']['first\_frame\_out\_f2'][1],predicted\_values\_first))\*\*0.5

RMSE\_model\_second = (mean\_squared\_error(data\_dict['valid']['second\_frame'][1],predicted\_values\_second))\*\*0.5

Вычисление среднего запаса и RMSE модели

def income(true\_target, pred\_target):

sort\_Series = pd.Series(pred\_target).sort\_values(ascending=False)[:FINAL\_MAX\_POINTS]

true\_target\_sort = (true\_target

.reset\_index(drop = True)[sort\_Series.index])

sum\_true = true\_target\_sort.sum()

return round((sum\_true \* PRICE\_PER\_BARREL) - BUDGET\_PER\_REGION,2)

Вычисление прибыли

def confidence\_interval(true\_target,pred\_target):

samples = []

for i in tqdm(range(10000)):

sample = pd.Series(pred\_target).sample(n = PRE\_MAX\_POINTS, replace=True, random\_state=state)

samples.append(income(true\_target,sample))

samples = pd.Series(samples)

print(samples.mean())

print(samples.apply(lambda x: x < 0).sum()/len(samples)\*100,"%")

lower = samples.quantile(0.025)

upper = samples.quantile(0.975)

return round(lower,2), round(upper,2)

Вычисление доверительного интервала

1. **Результаты**

В первом наборе данных признаки f0 и f1 отрицательно коррелируют относительно друг друга и f2 слабо положительно коррелирует с целевым признаком. Во втором очень сильно коррелирует целевой признак и f2. В третьем так же есть коррелирующие признаки, такие как f2 и product.

Если в случае с вторым набором все достаточно понятно, там очень высокая корреляция и признак f2 следует удалить, то вот в случае с другими двумя выборками стоит опираться на результат, который мы получим на моделях, следовательно необходимо подготовить 3 выборок:

1. Первая со всеми столбцами
2. Вторая без f2
3. Третья со всеми столбцами

В качестве результата для датасета из 3 регионов по 40 скважин получаем:

Средний запас предсказанного на выборке 1 сырья: 116.4882905615353 RMSE модели: 52.75846245311124

Средний запас предсказанного на выборке выборке 2 сырья: 60.896194431332006 (тыс.баррелей)

RMSE модели: 42.993319725878465

Средний запас предсказанного на выборке 3 сырья: 93.54679871989383 (тыс.баррелей)

RMSE модели линейной регрессии на выборке 3 : 40.404795778577686

Средняя прибыль в первой локации: 299715.04082699935 рублей

95% доверительный интервал: (135597.19, 445452.66)

Вероятность убытков: 0.02 %

Средняя прибыль во второй локации: 23380.939600000263 рублей

95% доверительный интервал: (-149628.34, 226369.65)

Вероятность убытков: 42.71 %

Средняя прибыль в третьей локации: 354192.9837779999 рублей

95% доверительный интервал: (221090.03, 489090.39)

Вероятность убытков: 0.0 %

1. **Вывод**

Лабораторная работа была направлена на определение наиболее прибыльного региона, учитывая параметры скважин из выборки и объем нефти. Для этого была использована модель

линейной регрессии, которая позволила определить зависимость переменной от других факторов.

После обучения модели на каждой выборке было выявлено, что самым прибыльным регионом является третий (354192 руб), при этом его среднеквадратичное отклонение на валидационной выборке было достаточно низким по сравнению с другими регионами.

Второй регион оказался наименее прибыльным.

Для каждого региона была проведена оценка рисков и прибыльности с помощью технологии bootstrap. Результаты соответствовали предсказанным значениям. Самым доходным регионом

остался последний, а наименее прибыльным – второй. Однако доверительный интервал для второго

региона был близок к убыточности. У первого и последнего риски составили – 0.02 и 0.00 соотвественно.

Таким образом, данная лабораторная работа позволила выявить наиболее прибыльный регион и оценить риски и прибыльность каждого из них.