**Описание проекта**

Вы работаете в IT-компании, которая выполняет на заказ проекты по машинному обучению.

К вам обратился фермер, владелец молочного хозяйства «Вольный луг». Он хочет купить бурёнок, чтобы расширить поголовье стада коров. Для этого он заключил выгодный контракт с ассоциацией пастбищ «ЭкоФерма».

Условия позволяют фермеру очень тщательно отобрать коров. Он определяет качество молока по строгой методике, и при этом ему нужно выполнять свой план развития молочного хозяйства. Фермер хочет, чтобы каждая бурёнка давала не менее 6000 килограммов молока в год, а её надой был вкусным — строго по его критериям, ничуть не хуже. А продавцы и технологи так и норовят приукрасить своих коровок!

Поэтому он просит вас разработать модель машинного обучения, которая поможет ему управлять рисками и принимать объективное решение о покупке. «ЭкоФерма» готова предоставить подробные данные о своих коровах.

**Вам нужно создать две прогнозные модели для отбора бурёнок в поголовье:**

1. Первая будет прогнозировать возможный удой коровы (целевой признак Удой);
2. Вторая — рассчитывать вероятность получить вкусное молоко от коровы (целевой признак Вкус молока).

С помощью модели нужно отобрать коров по двум критериям:

* средний удой за год — не менее 6000 килограммов;
* молоко должно быть вкусным.

**Описание данных**

Для студентов, начавших работу над проектом до 18:00 6 октября 2023 года

Вы будете работать с тремя датасетами:

[ferma\_main.csv](https://code.s3.yandex.net/datasets/ferma_main.csv)

[ferma\_dad.csv](https://code.s3.yandex.net/datasets/ferma_dad.csv)

[cow\_buy.csv](https://code.s3.yandex.net/datasets/cow_buy.csv)

Вы можете их скачать, нажав на название каждого.

1. Файл [ferma\_main.csv](https://code.s3.yandex.net/datasets/ferma_main.csv) содержит данные о стаде фермера на текущий момент. Описание данных:
   * id — уникальный идентификатор коровы.
   * Удой, кг — масса молока, которую корова даёт в год (в килограммах).
   * ЭКЕ (Энергетическая кормовая единица) — показатель питательности корма коровы.
   * Сырой протеин, г — содержание сырого протеина в корме (в граммах).
   * СПО (Сахаро-протеиновое соотношение) — отношение сахара к протеину в корме коровы.
   * Порода — порода коровы.
   * Тип пастбища — ландшафт лугов, на которых паслась корова.
   * порода папы\_быка — порода папы коровы.
   * Жирность,% — содержание жиров в молоке (в процентах).
   * Белок,% — содержание белков в молоке (в процентах).
   * Вкус молока — оценка вкуса по личным критериям фермера, бинарный признак (вкусно, не вкусно).
   * Возраст — возраст коровы, бинарный признак (менее\_2\_лет, более\_2\_лет).

Признаки в датасете ferma\_main.csv можно разделить на группы.

* + Характеристики коровы: id, Порода, порода папы\_быка, Возраст.
  + Характеристики корма: ЭКЕ (Энергетическая кормовая единица), Сырой протеин, г, СПО (Сахаро-протеиновое соотношение).
  + Характеристика пастбища: Тип пастбища.
  + Характеристики молока: Удой, кг, Жирность,%, Белок,%, Вкус молока.

Такое разделение поможет вам понять, как взаимосвязаны признаки в датасете.

1. Файл [ferma\_dad.csv](https://code.s3.yandex.net/datasets/ferma_dad.csv) хранит имя папы каждой коровы в стаде фермера. Описание данных:
   * id — уникальный идентификатор коровы.
   * Имя Папы — имя папы коровы.
2. Файл [cow\_buy.csv](https://code.s3.yandex.net/datasets/cow_buy.csv) — это данные о коровах «ЭкоФермы», которых фермер хочет изучить перед покупкой. Описание данных:
   * Порода — порода коровы.
   * Тип пастбища — ландшафт лугов, на которых паслась корова.
   * порода папы\_быка — порода папы коровы.
   * Имя\_папы — имя папы коровы.
   * Текущая\_жирность,% — содержание жиров в молоке (в процентах).
   * Текущий\_уровень\_белок,% — содержание белков в молоке (в процентах).
   * Возраст — возраст коровы, бинарный признак (менее\_2\_лет, более\_2\_лет).

Данные в файлах ferma\_main.csv и cow\_buy.csv похожи, но у датасета cow\_buy.csv несколько особенностей.

* + Содержание белков и жиров в молоке указано на момент продажи — когда «ЭкоФерма» давала коровам свой корм.
  + Параметры корма ЭКЕ (Энергетическая кормовая единица), Сырой протеин, г и СПО (Сахаро-протеиновое соотношение) отсутствуют. Технологи заказчика пересмотрели подход к кормлению: для новых коров планируется увеличить значения каждого из этих параметров на 5%.
  + Кроме того, отсутствуют признаки Удой, кг и Вкус молока. Это целевые признаки, именно их предстоит предсказать!

**Инструкция по выполнению проекта**

Для исследовательского анализа данных и оценки качества моделей вы можете создавать и использовать функции. Они заметно упростят вам работу:

* вам не придётся постоянно повторять однотипный код и обновлять данные;
* вы уменьшите риск допустить ошибку в коде;
* ваш код будет более лаконичным и понятным;
* вы сможете использовать ваши функции для будущих проектов.

При проверке этого проекта ревьюеры не будут негативно оценивать отсутствие функций — используйте их по желанию.

**Шаг 1. Загрузите и изучите данные**

1. Загрузите данные в датафреймы pandas. Файлы можно скачать по ссылкам в описании, они даже доступны через пути:
   * /datasets/ferma\_main.csv
   * /datasets/ferma\_dad.csv
   * /datasets/cow\_buy.csv

Для студентов, начавших работу над проектом до 18:00 6 октября 2023 года

1. Изучите данные. Определите, соответствуют ли они описанию задачи и изученным в прошлых спринтах стандартам оформления данных. При необходимости переименуйте столбцы.

**Шаг 2. Предобработка данных**

Проверьте данные на наличие пропусков и дубликатов. Узнайте, корректны ли типы данных у каждого признака. При необходимости устраните все проблемы с данными.

**Шаг 3. Исследовательский анализ данных**

1. Проведите статистический анализ всех признаков.
2. Постройте графики для каждого признака. Помните, что количественным и категориальным признакам подходят разные графики.
3. Сделайте выводы о данных:
   * напишите, заметили ли вы выбросы, необычное распределение количественных признаков или другие особенности;
   * расскажите, как вы планируете улучшить данные.

**Шаг 4. Проведите корреляционный анализ признаков в датасете** ferma\_main.csv

1. Изучите взаимосвязь между признаками: рассчитайте коэффициенты корреляции между всеми признаками.
2. Постройте диаграммы рассеяния scatterplot для признака Удой, кг и всех количественных признаков с учётом значения категориальных признаков. Пример: диаграмма рассеяния для Удой, кг и Сырой протеин, г, цвет точек диаграммы отображает признак Порода. Учитывайте, что виды связи в подгруппах могут различаться — например, для разных пород.
3. Сделайте выводы о взаимосвязи признаков.
   * Какой тип взаимосвязи между целевым и входными признаками?
   * Есть ли нелинейные связи между целевым и входными признаками? Нужно ли вводить в модель дополнительные признаки, чтобы это учесть?
   * Наблюдается ли мультиколлинеарность между входными признаками?

**Шаг 5. Задача регрессии**

Вам предстоит обучить три модели простой линейной регрессии.

1. Подготовьте данные датасета ferma\_main.csv и обучите на них первую модель LinearRegression.
   * Используйте в качестве целевого признак Удой, кг.
   * Отберите признаки, которые подходят для модели линейной регрессии, аргументируйте выбор. Опирайтесь не только на наличие или отсутствие линейной связи между признаками, но и на причинно-следственные связи между удоем коров и группами признаков из описания датасета.
   * Разбейте данные на тренировочную и тестовую выборки.
   * Кодируйте категориальные признаки с помощью OneHotEncoder.
   * Масштабируйте количественные признаки с помощью StandardScaler.
   * Обучите модель линейной регрессии на подготовленных данных.
   * Оцените качество модели на тестовой выборке: рассчитайте R², проанализируйте остатки с помощью графиков.
   * Оформите выводы, укажите:
     + значение R² модели;
     + какие выводы следуют из анализа остатков;
     + нужны ли дополнительные исследования.
2. Подготовьте данные с учётом связи входных признаков с целевым и обучите на них вторую модель LinearRegression.
   * Посмотрите на диаграммы рассеяния: на них можно увидеть, что некоторые входные признаки связаны с целевым нелинейно. Исключите нелинейность, чтобы эти признаки были информативны для линейной регрессии. Вот несколько способов, которыми это можно сделать:
     + Вы обнаружите, что СПО и Удой связаны нелинейно. На их диаграмме рассеяния наблюдения сгруппированы в два кластера, причём есть такое значение СПО, которое можно считать границей между кластерами. Чтобы устранить нелинейность, преобразуйте СПО в категориальный бинарный признак, использовав эту границу.
     + Вы обнаружите, что ЭКЕ и Удой также связаны нелинейно. Нелинейность между ними можно исключить, если ввести новый признак ЭКЕ в квадрате.
   * Повторите шаги из пункта 1 и создайте вторую модель: подготовьте данные, обучите модель, оцените её качество и напишите выводы.
3. Добавьте в данные новый признак и обучите на них третью модель LinearRegression.
   * Добавьте признак Имя Папы из файла ferma\_dad.csv, соединив его с остальными данными ferma\_main.csv по ключу.
   * Устраните нелинейность между признаками.
   * Повторите шаги из пункта 1 и создайте третью модель: подготовьте данные, обучите модель, оцените её качество и напишите выводы.
4. Сравните качество трёх моделей линейной регрессии, сопоставьте их R² и остатки.
5. Оцените качество каждой модели по трём метрикам: MSE, MAE, RMSE. Используйте функции для расчёта метрик.
   * Выберите лучшую модель;
   * Укажите доверительный интервал прогноза лучшей модели. Для этого рассчитайте квантили ошибки модели на тренировочных данных.
6. С помощью лучшей модели спрогнозируйте удой коров, которых фермер хочет купить. Данные о них находятся в датасете cow\_buy.csv.
   * Добавьте недостающие признаки ЭКЕ (Энергетическая кормовая единица), Сырой протеин, г и СПО (Сахаро-протеиновое соотношение): возьмите из датасета ferma\_main.csv средние значения каждого из признаков и увеличьте их на 5%.
   * Выполните подготовку данных, аналогичную той, которую делали для тренировочной выборки.
   * Получите прогноз удоя коров.

**Шаг 6. Задача классификации**

1. Подготовьте данные датасета ferma\_main.csv и обучите на них модель LogisticRegression.
   * Используйте в качестве целевого признак Вкус молока.
   * При необходимости преобразуйте имеющиеся или добавьте новые признаки.
   * Разбейте данные на тренировочную и тестовую выборки.
   * Кодируйте категориальные признаки с помощью OneHotEncoder.
   * Масштабируйте количественные признаки с помощью StandardScaler.
   * Обучите модель логистической регрессии на подготовленных данных.
   * Оцените качество модели на тестовой выборке: рассчитайте метрики accuracy, recall, precision.
   * Постройте матрицу ошибок.
2. Проанализируйте модель логистической регрессии.
   * Объясните, какая ошибка критичнее всего при покупке коров: первого или второго рода. Помните, что главное для заказчика — вкусное молоко. Сформулируйте ответ с учётом метрик recall и precision, выберите более важную из них.
   * Сведите критичную ошибку к нулю: для этого измените порог принадлежности к классам. Ещё раз рассчитайте значения метрик recall и precision. Какое значение порога вы выбрали? Как изменились значения метрик?
   * Оформите выводы: скажите, что нужно сделать дальше, исходя из анализа модели?
3. С помощью обученной модели спрогнозируйте вкус молока коров, которых фермер хочет купить. Данные о них находятся в датасете cow\_buy.csv.
   * Выполните подготовку данных, аналогичную той, которую делали для тренировочной выборки.
   * Получите прогноз вкуса молока коров.

**Шаг 7. Сделайте итоговые выводы**

* Добавьте прогнозы вкуса молока и удоя коров «ЭкоФермы» в датафрейм с коровами, покупку которых рассматривает фермер.
* Отберите коров, для которых спрогнозирован удой более 6000 килограммов в год и вкусное молоко.
* Коротко опишите результаты работы моделей для фермера.
  + Сколько коров он может купить из стада на продажу с минимальным риском для себя? Почему?
  + Какие ещё рекомендации можно ему дать?
* Сделайте выводы о моделях линейной регрессии и логистической регрессии.
  + Какие способы улучшения качества регрессионной модели вы использовали в этой задаче?
  + Возможно ли ещё улучшить результаты регрессионной модели? Если можно, то как? Если нет, то почему?
  + Какие метрики моделей классификации, в том числе логистической регрессии, важны при анализе рисков или экономических расчётов?

**Как будут проверять мой проект**

Мы подготовили критерии оценки проекта, которыми руководствуются ревьюеры. Прежде чем приступить к решению кейса, внимательно их изучите.

На что обращают внимание ревьюеры, проверяя проект:

* как вы готовите данные к обучению, все ли особенности увидели и обрабатываете;
* хорошо ли объясняете шаги по предобработке данных и исследовательскому анализу;
* как проводите корреляционный анализ, проводите ли исследования в подгруппах;
* используете ли при корреляционном анализе графики;
* какие выводы об исследовании задачи делаете;
* корректно ли разбиваете данные на выборки;
* верно ли выбираете методы моделирования;
* корректно ли проводите обучение;
* корректно ли оцениваете модель по метрикам;
* корректно ли исследуете остатки модели;
* следите за структурой проекта и поддерживаете аккуратность кода;
* правильно ли используете модели;
* соблюдаете ли вы требования заказчика, даёте ли ему практические рекомендации.

<https://gitverse.ru/maxtiandi/training_projects/content/main/farm_risk_management>

# Управление рисками и принятие решения о покупке коров хозяйства "Вольный луг"

К IT-компании, которая выполняет на заказ проекты по машинному обучению, обратился фермер, владелец молочного хозяйства «Вольный луг». Он хочет купить бурёнок, чтобы расширить поголовье стада коров. Для этого он заключил выгодный контракт с ассоциацией пастбищ «ЭкоФерма». Условия позволяют фермеру очень тщательно отобрать коров. Он определяет качество молока по строгой методике, и при этом ему нужно выполнять свой план развития молочного хозяйства.

**Цель исследования:** Разработать модель машинного обучения, которая поможет ему управлять рисками и принимать объективное решение о покупке.

**Задачи исследования:**

1. Создать две прогнозные модели для отбора бурёнок в поголовье:
   * Первая будет прогнозировать возможный удой коровы (целевой признак Удой);
   * Вторая — рассчитывать вероятность получить вкусное молоко от коровы (целевой признак Вкус молока).
2. С помощью модели нужно отобрать коров по двум критериям:
   * средний удой за год — не менее 6000 килограммов;
   * молоко должно быть вкусным.

**Описание данных:**

1. Датасэт №1 - *ferma\_main.csv* - содержит данные о стаде фермера на текущий момент
   * id — уникальный идентификатор коровы.
   * Удой, кг — масса молока, которую корова даёт в год (в килограммах).
   * ЭКЕ (Энергетическая кормовая единица) — показатель питательности корма коровы.
   * Сырой протеин, г — содержание сырого протеина в корме (в граммах).
   * СПО (Сахаро-протеиновое соотношение) — отношение сахара к протеину в корме коровы.
   * Порода — порода коровы.
   * Тип пастбища — ландшафт лугов, на которых паслась корова.
   * порода папы\_быка — порода папы коровы.
   * Жирность,% — содержание жиров в молоке (в процентах).
   * Белок,% — содержание белков в молоке (в процентах).
   * Вкус молока — оценка вкуса по личным критериям фермера, бинарный признак (вкусно, не вкусно).
   * Возраст — возраст коровы, бинарный признак (менее\_2\_лет, более\_2\_лет).
2. Датасет №2 *ferma\_dad.csv* - хранит имя папы каждой коровы в стаде фермера.
   * id — уникальный идентификатор коровы.
   * Имя Папы — имя папы коровы.
3. Датасет №3 *cow\_buy.csv*  - это данные о коровах «ЭкоФермы», которых фермер хочет изучить перед покупкой.
   * Порода — порода коровы.
   * Тип пастбища — ландшафт лугов, на которых паслась корова.
   * порода папы\_быка — порода папы коровы.
   * Имя\_папы — имя папы коровы.
   * Текущая\_жирность,% — содержание жиров в молоке (в процентах).
   * Текущий\_уровень\_белок,% — содержание белков в молоке (в процентах).
   * Возраст — возраст коровы, бинарный признак (менее\_2\_лет, более\_2\_лет).

# Вывод

**В ходе работы над проектом была разработана модель машинного обучения, которая сможет помочь заказчику управлять рисками и принимать объективное решение о покупке коров на ферму. Рекомендуемые коровы к покупке:**

| **города** | **тип\_пастбища** | **порода\_папы** | **имя\_папы** | **текущая\_жирность** | **текущий\_белок** | **возраст** | **ожидаемый\_удой** | **ожидаемый\_вкус\_молока** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Вис Бик Айдиал | Холмистое | Айдиал | Геркулес | 3.58 | 3.076 | более 2 лет | 6755 | вкусно |
| Вис Бик Айдиал | Холмистое | Айдиал | Геркулес | 3.58 | 3.076 | более 2 лет | 6755 | вкусно |
| РефлешнСоверинг | Холмистое | Айдиал | Буйный | 3.40 | 3.079 | менее 2 лет | 6501 | вкусно |
| Вис Бик Айдиал | Холмистое | Соверин | Соловчик | 3.73 | 3.079 | более 2 лет | 6084 | вкусно |

Исследование было разделено на несколько этапов:

**1. Загрузка и подготовка данных**

* Импортировали нужные библиотеки
* ferma\_main
  + Обнаружили 5 явных дубликатов
  + Определили, что стоит изменить названия столбцов и привести их к единому виду
  + Не выявили пропущенных значений
  + Определили, что с типами данных все в порядке
* ferma\_dad
  + Определили, что стоит изменить названия столбцов и привести их к единому виду
  + В остальном данные оказались в порядке и пригодными для использования
* cow\_buy
  + Обнаружили 4 явных дубликата -> решили не убирать, поскольку данных в датасете мало, как и признаков. Вполне вероятно, что данные и признаки сходятся для разных коров
  + Определили, что стоит изменить названия столбцов и привести их к единому виду
  + Пропущенных значений не было выявлено
  + Определили, что с типами данных все в порядке

**2. Предобработка данных**

* Названия столбцов привели к единому, удобному виду
* Убрали 2 неявных дубликата из датасета ferma\_main
* Убрали 5 явных дубликатов из ferma\_main

**3. Исследовательский анализ данных**

* Выявили, что почти все признаки распределены нормально и не требуют особых корректировок выбросов
* Убрали аномальное значение признака [удой]**== 45616**
* В признаке [белок] избавились от выбросов -> перезаписали датафрейм с условием, в которое входят данные из диапазона:
  + < 3.069
  + > 3.081
* В распределение категориальных переменных критических изменений внесено не было

**4. Корреляционный анализ данных**

* эке: Выявили положительную взаимосвязь выше умеренного значения. С увеличением "удоя" постепенно увеличивается и эке. По графикам распределения, корреляция справедлива для двух пород.
* сырой\_протеин: Взаимосвязь между содержанием сырого протеина и удоем не так ясна. На графике диаграмме рассеяния виден довольно большой разброс значений, однако наблюдается общий тренд на прямую пропорциональную зависимость. Связь умеренно положительная.
* спо: Выявили, что СПО зависит от целевого признака - коэффициент корреляции выше среднего. В особенности видна резкая тенденция возвышения значений после - определен "порог перехода" в значении спо == 0.91
* жирность: Выявлена умеренно положительная зависимость. Более всего связь раскрывается на переходе значений жирность == 3.6 и 3.7.
* белок: Связь между этим признаком и целевым - нелинейная. Но стоит отметить, что наблюдается зависимость между признаками белок и порода. На меньших значения "белка" - больше значений с породой **РефлешнСоверинг**, а на бОльших значениях - больше значений породы Вис Бик Айдиал.
* возраст имеет сильнейшую взаимосвязь с целевым признаком удой. Коэффициент взаимосвязи: *1.00*
* Между целевым и входными признаками связь в основном **линейная**, но наличие нелинейных связей также возможно для отдельных пород.
* Яркие нелинейные связи не выявлены, в связи с чем **добавление новых признаков - не потребовалось**
* **Что касается мультиколлинеарности:** матрица показала, что между некоторыми входными признаками есть умеренная корреляция, но нет явных признаков сильной мультиколлинеарности, с которой стоило бы справляться.

**5. Решение задачи регресии (предсказание удоя)**

* Обучили 3 модели линейной регрессии и проанализировали их остатки
* Выявили лучшую модель из трех на валидационной выборке, показатели которой:
  + MAE - 146.4625952324208
  + MSE - 37967.73485485037
  + RMSE - 194.85311097041887
  + **95% доверительный интервал** = (6070.497027590764, 6211.5065511559)
* Собрали train + valid выборки в единую обучающую и протестировали модель на тестовой выборке:
  + Получили предсказания по всем коровам продавца
  + Добавили столбец с предсказаниями в таблицу продавца

**6. Решение задачи классификации (предсказания вкуса молока)**

* Подобрали параметры и обучили модель логистической регрессии
* Проанализировали модель и минимизировали количество критических ошибок первого рода с 42 до 9, подобрав новый порог для вероятности предсказаний
* Определили 4-х коров, чье молоко предсказано как "вкусное"
* Отобрали для рекомендации 4 коровы, учитывая оба условия заказчика