1. Яндекс.Музыка	Первый проект для отработки базовых инструментов библиотеки Pandas. Цель исследования — проверить ряд гипотез на данных о пользователях Яндекс.Музыки в Москве и Санкт-Петербурге, предварительно проведя обработку данных.
2. Исследование надежности заемщиков	Задача проекта: выяснить, какие факторы (семейное положение, наличие детей) влияют на факт возврата банковского кредита в срок. Был проведён обзор данных, их обработка (заполнены пропуски, изменены и удалены экстремальные значения), созданы словари, проведена категоризация клиентов по уровню дохода и по целям кредита. В качестве основного исследовательского инструмента использовались сводные таблицы. В работе использовалась библиотека Pandas.
3. Исследование объявлений о продаже квартир	По представленным данным сервиса Яндекс Недвижимость необходимо выявить параметры, которые влияют на цену недвижимости. В ходе исследования была проведена предобработка данных – в учебных целях особо тщательно (с заменой экстремальных значений с помощью функций). Были созданы дополнительные столбцы в исходной таблице, проведён исследовательский анализ данных. Для исследования использовалась библиотека Pandas.
4. Оператор связи - проверка гипотез	Цель исследования состоит в изучении данных о том, как абоненты оператора мобильной связи, которые подключены к одному из двух тарифов, используют свой тариф. Нужно выяснить, какой из них приносит оператору больше выручки. Исследование включало проверку гипотез о различии средней выручки у разных групп абонентов. Использовались библиотеки Pandas, Scipy, Numpy, Seaborn, Matplotlib.
5. Исследование продаж компьютерных игр	Представлены данные по продажам игр в различных жанрах, для различных платформ в нескольких регионах мира (таблица содержит 11 столбцов, 16,7 тыс. строк). Цель проекта - выявить закономерности, определяющие коммерческую успешность игры. Для исследования использовались библиотеки Pandas, Numpy, Matplotlib, Seaborn, Scipy.
6. Рекомендация тарифов мобильной связи	По данным о поведении клиентов мобильного оператора, которые перешли на тарифы "Ультра" и "Смарт", нужно построить модель для задачи классификации (на несбалансированной выборке), которая выберет наиболее подходящий из этих двух тариф для пользователей, которые сейчас подключены к архивным тарифам, - на основе параметров использования мобильной связи. Предобработка данных не проводится. Использовались библиотеки Pandas, Seaborn, Scikit-Learn.
7. Прогнозирование оттока клиентов банка	По представленным данным о клиентах «Бета-Банка» и их поведении нужно спрогнозировать, уйдёт клиент из банка в ближайшее время или нет.

	Задача состоит в анализе различных моделей и выборе подходящей - с предельно большим значением F1-меры. Работа проводилась с помощью библиотек Pandas, Matplotlib, Seaborn, Scikit-learn.
8. Выбор локации для скважины	Целью исследования является выбор региона для разработки нефтяных месторождений - из трёх регионов, по которым известны параметры 10000 скважин. Необходимо с помощью модели ML определить регион, в котором добыча нефти из 200 скважин принесёт наибольшую прибыль. Однако известно, что компания может исследовать только 500 случайных скважин в регионе. Для выполнения задач использовались библиотеки Pandas, Seaborn, Scikit-learn, Numpy, Scipy.
9. Защита персональных данных клиентов страховой компании	Цель проекта - защитить данные клиентов страховой компании «Хоть потоп». Для этого нужно разработать такой метод преобразования данных, чтобы по ним было сложно восстановить персональную информацию. Нужно защитить данные, чтобы при преобразовании качество моделей машинного обучения не ухудшилось. Подбирать наилучшую модель не требуется. Для выполнения задач использовались библиотеки Pandas, Numpy, Scikit-learn.
10. Определение стоимости автомобилей	Сервис по продаже автомобилей с пробегом «Не бит, не крашен» разрабатывает приложение для привлечения новых клиентов. В нём можно быстро узнать рыночную стоимость своего автомобиля. По историческим данным: техническим характеристикам, комплектации и ценам автомобилей – нужно построить модель для определения стоимости. Задача: выбрать модель по качеству предсказаний, скорости предсказания и времени обучения. Для выполнения задач использовались библиотеки Pandas, Numpy, Seaborn, Scikit-learn, LightGBM, CatBoost.
11. Предсказание спроса на такси	Отработка предсказания на временных рядах. Компания «Чётенькое такси» собрала исторические данные о заказах такси в аэропортах. Чтобы привлекать больше водителей в период пиковой нагрузки, нужно спрогнозировать количество заказов такси на следующий час. Цель исследования: обучить модели с различными гиперпараметрами, чтобы спрогнозировать количество заказов такси на следующий час. Значение метрики RMSE должно быть не более 48. Для выполнения задач использовались библиотеки Pandas, Statsmodels, Matplotlib, Numpy, Scikit-learn, XGBoost, CatBoost.
12. Проект для "Викишоп" с BERT	Интернет-магазин "Викишоп" запускает сервис, предоставляющий пользователям публиковать и дополнять описания товаров, комментировать изменения других пользователей. Цель проекта - разработать модель для поиска токсичных комментариев, чтобы можно было отправлять их на модерацию, качество определения токсичных комментариев должно быть не менее 0,75 по F1.

	Решено было попробовать достичь цели двумя способами: с помощью BERT и с помощью TF-IDF. Использовались следующие библиотеки: Pandas, Numpy, Natural Language Toolkit, Scikit-learn, XGBoost, Pytorch, Transformers.
13. Определение возраста покупателей	Проект посвящён обработке фотографий с помощью компьютерного зрения. Сетевой супермаркет «Хлеб-Соль» внедряет систему компьютерного зрения для обработки фотографий покупателей. Задача - определить возраст покупателей. Для выполнения этой задачи необходимо построить модель, которая будет по фотографии определять приблизительный возраст человека с качеством около 7 по метрике MAE. Для решения задачи использовались библиотеки Pandas, Numpy, Matplotlib, Tensorflow. Решено было выбрать нейросеть ResNet50, однако с "кастомизированными" верхними слоями.
14. Прогнозирование оттока клиентов оператора связи	В общем виде цель поставлена таким образом: оператор связи хочет прогнозировать отток клиентов для того, чтобы предлагать специальные условия клиентам из "группы риска". Первичная задача состоит в анализе данных и составлении плана работы на последующие этапы. После уточнения данных была сформулирована основная задача работы: предложить модель бинарной классификации (планирует ли клиент покинуть компанию в ближайшее время или нет), которая дала бы качество свыше 0.85 по метрике ROC-AUC. Для решения задач использовались библиотеки Pandas, Matplotlib, Numpy, Seaborn, Phik, Scikit-learn, LightGBM, XGBoost, CatBoost.