Проекты по Classic ML и Data Science | ML course FAMCS BSU

Название проекта	Описание	Источник данных	Методы	Таргет (y)	Что нужно сделать	Ключевые метрики (ориентиры)	Ожидаемые выводы / инсайты
Прогнозирование цен на недвижимость	Построение моделей регрессии для предсказания стоимости жилья. Включает обработку выбросов, инжиниринг признаков, сравнение моделей.	California Housing Boston Housing Russian Housing	Linear Regression, Ridge, Lasso, Decision Trees, Random Forest, Gradient Boosting	Стоимость объекта в долларах/рублях (непрерывная)	- Очистка данных, устранение выбросов - Feature engineering (площадь, год постройки, локация и др.) - Обучение и сравнение линейных и нелинейных регрессионных моделей - Подбор гиперпараметров, кросс-валидация	RMSE ≈ 35–45 тыс.$, MAE < 30 тыс.$, R^2 > 0,80	Определить, какие признаки сильнее всего влияют на цену и насколько нелинейные модели превосходят линейные
Классификация заболеваний	Бинарная/многоклассовая классификация заболеваний на основе медицинских показателей. Работа с несбалансированными данными, важность интерпретации.	Pima Indians Diabetes Heart Disease UCI Breast Cancer Wisconsin	Logistic Regression, SVM, Random Forest, XGBoost	Наличие/тип заболевания (0/1 или 0 … K)	- Работа с несбалансированными данными (re-/under-sampling, class weights) - Обучение базовых и ансамблевых классификаторов - Интерпретация важности признаков	ROC-AUC ≥ 0,85, Recall ≥ 0,80, PR-AUC ≥ 0,80	Какие медицинские показатели наиболее информативны; цена ошибки FN vs FP
Анализ оттока клиентов	Предсказание клиентов, которые могут уйти из компании. Поиск ключевых факторов, балансировка классов, оптимизация порога классификации.	Telco Customer Churn Bank Customer Churn	Различные классификаторы, техники сэмплирования, анализ важности признаков	Флаг ухода клиента (binary)	- Создание поведенческих и демографических признаков - Балансировка классов, оптимизация порога - Выделение top-N клиентов по риску	ROC-AUC 0,75 -- 0,85, F1 0,55 -- 0,65, Precision@10 % ≥ 0,30	Факторы, повышающие риск ухода; потенциальная экономия от удерживающих кампаний
Кластеризация и сегментация клиентов	Выявление скрытых паттернов поведения и сегментация клиентов. Определение оптимального числа кластеров, интерпретация результатов, визуализация.	Online Retail II Customer Segmentation	K-means, Hierarchical Clustering, DBSCAN, Gaussian Mixture Models	-- (unsupervised)	- Построить RFM/поведенческий признак-пространство - Подобрать оптимальное K (Elbow, Silhouette > 0,40) - Визуализировать кластеры, описать сегменты	Silhouette > 0,40, Calinski-Harabasz ↑	Получить 4-6 осмысленных сегментов, сформулировать для них маркетинговые гипотезы
Обнаружение аномалий в показателях датчиков	Выявление аномальных показаний в больших массивах сенсорных данных. Работа с мультивариативными временными рядами, дисбаланс классов.	Industrial IoT datasets NASA Bearing Dataset Sensor Fault Detection	One-class SVM, Isolation Forest, LOF, статистические методы	Метка "аномалия" (0/1) или скор	- Предобработка и нормализация временных рядов - Обучение One-class SVM / Isolation Forest / LOF - Настройка порога детектирования	PR-AUC ≥ 0,60, Recall ≥ 0,90 при FPR ≤ 0,05	Выявить ранние признаки отказа оборудования; снизить незапланированные простои
Прогнозирование спроса на товары/услуги	Прогнозирование спроса с учетом сезонных факторов и внешних переменных. Работа с многомерными данными, сложные зависимости.	Kaggle Store Item Demand Store Sales Forecasting	Ансамблевые модели, мета-прогнозирование, оптимизация гиперпараметров	Количество продаж (шт./день)	- Анализ сезонности и тренда, encoding календарных признаков - Обучение ансамблевых моделей/GBRT/MLP на лаговых фичах - Пост-обработка (moving avg, quantile adj.)	MAPE 10 -- 20 %, RMSE ↓ vs наивный прогноз минимум 30 %	Выделить SKU/магазины с наибольшей волатильностью; рекомендовать safety stock
Построение скоринговых моделей	Предсказание вероятности дефолта, построение скоринговой карты. Калибровка вероятностей, интерпретируемость, работа с дисбалансом.	Lending Club Loan Data Credit Risk datasets	Логистическая регрессия с регуляризацией, интерпретируемые модели градиентного бустинга	Вероятность дефолта (PD, 0-1)	- Кодирование категориальных, WOE/IV анализ - Логистическая регрессия + бустинг, калибровка - Разработка скор-карты	Gini > 40 %, KS > 0,35, Brier ≤ 0,18	Выявить кредитно-значимые факторы; предложить cut-off для разных risk-политик
Оптимизация промышленных процессов	Оптимизация параметров процессов для снижения энергопотребления или отходов. Многоцелевая оптимизация, работа с высокоразмерными данными.	Industrial Production Data Steel Industry Energy Consumption	Регрессионные модели, ансамблевые методы, алгоритмы оптимизации	KPI процесса (энергия, выход %)	- Построить регрессию KPI от параметров процесса - Использовать SHAP/Permutation для интерпретации - (Опц.) применить Bayesian Optimization	R^2 > 0,70, снижение KPI-cost на 5-10 %	Какие параметры влияют сильнее всего и до каких значений их стоит оптимизировать
Анализ и предсказание преступности	Выявление факторов и географических паттернов преступности, прогнозирование. Пространственный анализ, корреляции с социально-экономическими	Communities and Crime Chicago Crime Dataset	Пространственная регрессия, деревья решений, градиентный бустинг	Кол-во преступлений или вероятность события в районе	- Слияние данных о преступлениях с соц-экон показателями и ГИС - Пространственная регрессия / градиентный бустинг - Визуализация	RMSE ↓ 30 % vs mean baseline или ROC-AUC ≥ 0,80 (класс. задача)	Горячие зоны преступности; связь с уровнем безработицы, освещённостью и др.