ML & BD

План (очень примерно) на 7 занятий

1. **Основные термины и понятия**

Основные термины и понятия. Виды задач в ML: с учителем (классификация, регрессия, ранжирование), без учителя (кластеризация, понижение размерности, поиск аномалий), с частичным привлечением учителя, с подкреплением… Типичные примеры. Обзор метрик качества. Бизнес-метрики. Онлайн-метрики. Офлайн-метрики. Для классификации: матрица ошибок, accuracy, recall, precision, f-мера. Для регрессий: MSE, MAE, MAPE, AMAPE... Функция потерь. L2, L1, Huber Loss. Обобщающая способность. Переобучение, недообучение. Bias, Variance, Noise. Декомпозиция bias-variance. Обучение и оценивание алгоритма. Метрики обучения. AUC, Gini, log-loss. Пороги отсечения вероятности. Параметры и гиперпараметры. Разбиение датасета для обучения и оценки. Подбор гиперпараметров: стохастический градиентный спуск, генетические алгоритмы, перебор по сетке, случайный перебор. Отложенная выборка. Кросс-валидация: k-Fold, LOOCV. Лик.

2. **EDA и предварительная обработка данных.**

EDA - Разведочный анализ данных. Графическое представление датасета. Визуализация. Обзор основных типов графиков: line chart, scatter-plot, histogram, матрица парных корреляций… Библиотеки matplotlib, seaborn, plotly. Типичные проблемы и способы борьбы с ними. Аномальные значения и выбросы. Винзоризация. Шум. Пропущенные переменные. Фиктивная подстановка. Аппроксимации интерполяцией и определение наиболее вероятного значения. Категориальные признаки. Дублированные значения. Противоречивые данные. Ошибки в данных, их логическая и физическая неадекватность, погрешность измерения. Мультиколлинеарность. Несбалансированность. Стандартизация. Нормализация. Преобразование Бокса-Кокса.

3. **Линейные методы**

Линейная регрессия. Постановка задачи. Метод наименьших квадратов. Виды: line-line, line-log, log-line, log-log. Графическое представление. Дамми переменные сдвига и наклона. Интерпретация коэффициентов. Метрики качества R2, SSE и информационные критерии AIC, BIC, HQIC. Проблемы неправильной спецификации. Omitted Variable Bias. Логистическая регрессия. Интерпретация коэффициентов. «Штраф» за сложность. Регуляризация. Временные ряды. Модели тренда и сезонности. Статистические модели временных рядов. Авторегрессия. Адаптивные и модели. EWMA. Корреляции между временными рядами. Ложная корреляция. Отсутствие причинно-следственных связей. Двусторонних эффект. Стационарность. Взвешенное скользящее среднее.

4. **Метрические методы**

Метод опорных векторов. Разделяющая гиперплоскость. Опорные вектора. Случаи линейно разделимой и неразделимых выборок. Ширина разделяющей полосы. Функционал ошибки и его регуляризация. Ядра и спрямляющие пространства. Алгоритм k (взвешенных) ближайших соседей. Проклятие размерности. Масштабирование: нормализация, стандартизация. Евклидова и другие метрики расстояния. Случай категориальных переменных.

5. **Логические методы**

Решающие пни, деревья и леса. Методы расщепления. Критерии информативности для классификации: энтропия, Джини. Недостатки жадной стратегии и способы их устранения. Критерии остановки обучения. Отсечение ветвей. Ансамбли алгоритмов. Теорема Кондорсе о присяжных. Мудрость толпы. Бустинг. Оптимизация. Градиенты. Выбор величины шага. Выбивание из локальных минимумов. Ранняя остановка.

6. **Кластеризация**

Иерархические (восходящие, нисходящие) и плоские методы. Дендрограммы. Четкие и нечеткие. Способы объединения кластеров (одиночная, полная связь, взвешенные и центроидные методы). K-Means, на основе плотности – DBSCAN, на основе распределений – EM алгоритм, MeanShift. Визуализация и t-SNE. Метрики качества кластеризации.

7. **Отбор признаков, значимость и применимость в бизнесе**

Если заказчик сомневается или «сопротивляется»: некоторые лайфхаки для общения, объяснения и внедрения решений в продакшн. Определение «важности» признаков. Ridge, Lasso. Библиотеки Lime, Shap. Фильтрация на основе ранжирования по релевантности. Оберточные методы (forward / backward). Создание агрегированных признаков. Понижение размерности, метод главных компонент. А/Б тесты. Эксперименты.