Ключевые понятия и алгоритмы машинного обучения

- Алгоритм предобработки данных и их первичного анализа
  1. Проверка типов
  2. Проверка пропущенных значений
  3. Анализ распределений признаков по одному
     a. количественные: гистограмма\плотность, ящик с усами
     b. качественные: столбиковая диаграмма)
  4. Выявление аномалий
  5. Анализ взаимодействий признаков
     a. количественная-количественная: аналитический метод - корреляции, графический - скатерплот.
     b. количественная- качественные: аналитически метод - проверка гипотез (параметрический: стьюдент, непараметрический: маннауитни), графический - боксплот.
     c. качественные - качественные: аналитически метод – кф согласия Хи-квадрат, графический - круговая диаграмма

  *параметрические тесты предполагают, что параметры распределения ГС известны, непараметрические – неизвестны (менее точные)
  6. Анализ целевой переменной
  7. Анализ взаимодействия между признаками и целевой переменной

- Алгоритм нахождения модели:
  1. Изучаем проблематику бизнеса и самой темы (пытаемся понять суть явления)
  2. Генерируем разные признаки - любые.
  3. Проводим эксперимент:
     a. определяем тип модели и соответствующие ей способы подготовки данных
     b. отбираем признаки
     c. оптимизируем гиперпараметры модели
     d. анализируем метрики, значимости признаков

- Классификация моделей машинного обучения
  1. Линейные
  2. Деревянные
  3. Метрические

- Ансамбли моделей
  1. Бэггинг
  2. Лес (бэггинг на различных подпространствах признаков)
  3. Бустинг (каждая последующая модель учитывает ошибки предыдущей)
  4. Стекинг (использование ооф предсказаний базовых моделей в качестве признаков))

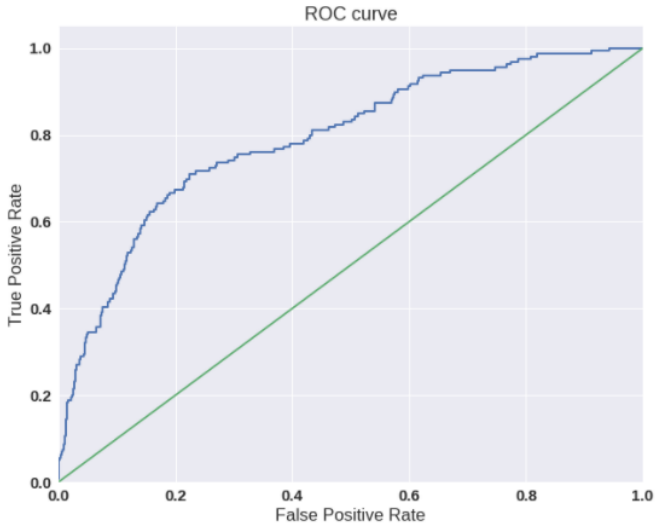
- Метрики классификации:

	$y = 1$	$y = 0$
$\hat{y} = 1$	True Positive (TP)	False Positive (FP)
$\hat{y} = 0$	False Negative (FN)	True Negative (TN)

1. Accuracy – доля правильных ответов
2. Precision – точность (ошибка 1-го рода) $\frac{tp}{tp+fp}$
3. Recall – полнота (ошибка 2-го рода) $\frac{tp}{tp+fn}$
4. F-score $\frac{precision*recall}{\beta^2*precision+recall} * (1+\beta^2)$
5. Roc-auc

$TPR = \frac{tp}{tp+fp}$ (точность)

$FPR = \frac{fp}{fp+tn}$ (какую долю из объектов $negative$ класса алгоритм предсказал неверно)



В идеальном случае, когда классификатор не делает ошибок (FPR = 0, TPR = 1) мы получим площадь под кривой, равную единице; в противном случае, когда классификатор случайно выдает вероятности классов, AUC-ROC будет стремиться к 0.5, так как классификатор будет выдавать одинаковое количество TP и FP.
Каждая точка на графике соответствует выбору некоторого порога. Площадь под кривой в данном случае показывает качество алгоритма (больше — лучше), кроме этого, важной является крутизна самой кривой — мы хотим максимизировать TPR, минимизируя FPR, а значит, наша кривая в идеале должна стремиться к точке (0,1).

- Метрики регрессии:

1. MAE
2. MSE
3. R2
4. MAPE