Алгоритмы построения моделей машинного обучения

Ксемидов Борис Сергеевич

19 марта, 2020

Метод k ближайших соседей

Метрический алгоритм для классификации объектов или регрессии.

Идея

Объекту присваивается тот класс, который является наиболее распространённым среди k соседей данного объекта.

Метод k ближайших соседей

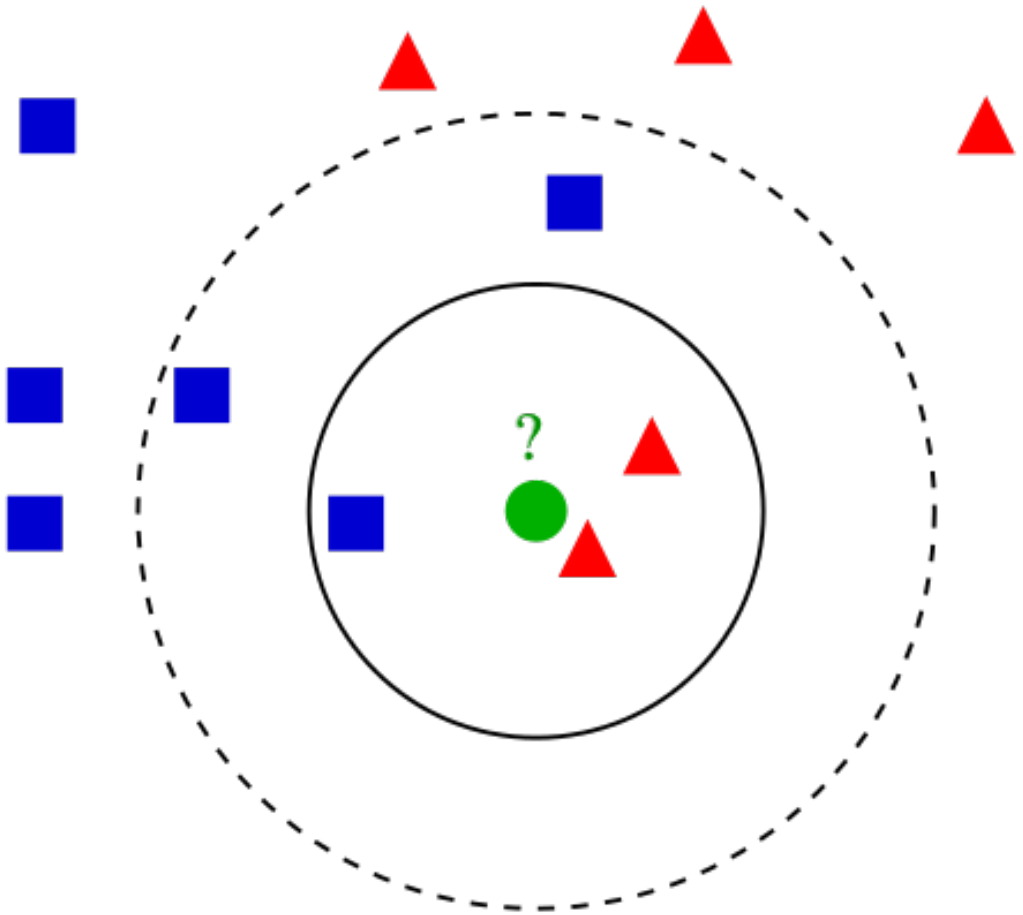

Рис.: Пространство признаков

Метод k ближайших соседей

Примеры метрик расстояния (w_j - вес для j-го соседа):

- $d(u, x_i) = (\sum_{j=1}^{n} w_j |f_j(u) - f_j(x_i)|^p)^{\frac{1}{p}}$ - метрика Минковского
- $d(u, x_i) = \sqrt{\sum_{j=1}^{n} w_j (f_j(u) - f_j(x_i))^2}$ - евклидова метрика

Решающие деревья

Логический алгоритм для классификации объектов или регрессии.

Идея

k-ичное дерево с решающими правилами в нелистовых вершинах (узлах) и некотором заключении о целевой функции в листовых вершинах (прогнозом).

Решающие деревья

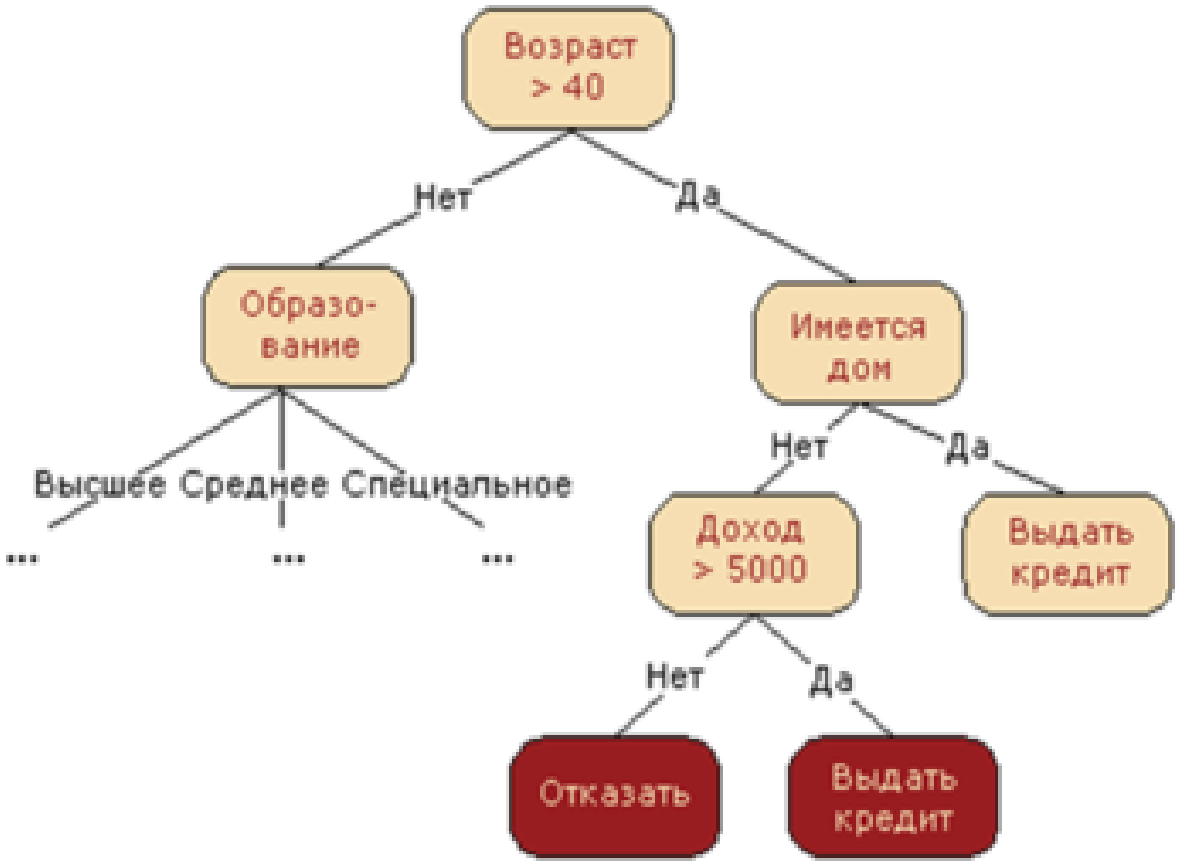

Рис.: Визуализация решающего дерева

Случайный лес

Идея

Использование большого ансамбля решающих деревьев, каждое из которых само по себе даёт очень невысокое качество классификации, но за счёт их большого количества результат получается хорошим.

Пример задачи

15 апреля 1912 года, во время первого плавания, широко известный «непотопляемый» Титаник затонул после столкновения с айсбергом. К сожалению, на борту не было достаточно спасательных шлюпок для всех, что привело к гибели 1502 из 2224 пассажиров и членов экипажа.

Цель

Необходимо построить прогностическую модель, которая отвечает на вопрос: «какие люди выжили с большей вероятностью?» используя данные о пассажирах (например, имя, возраст, пол, социально-экономический класс и т. д.).

Описание признаков

- PassengerId - уникальный идентификатор
- Survived - выжил или нет (1 или 0)
- Pclass - класс билета
- Name - имя
- Sex - пол
- Age - возраст
- SibSp - количество братьев, сестёр и супругов, присутствующих на Титанике
- Parch - количество родителей, детей, присутствующих на Титанике

Пример данных

Pass-eng-erId	Survived	Pclass	Name	Sex	Age	SibSp	Parch
1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0
2	1	1	Cumings, Mrs. John Bradley	female	38.0	1	0
3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0

Разделение по полу

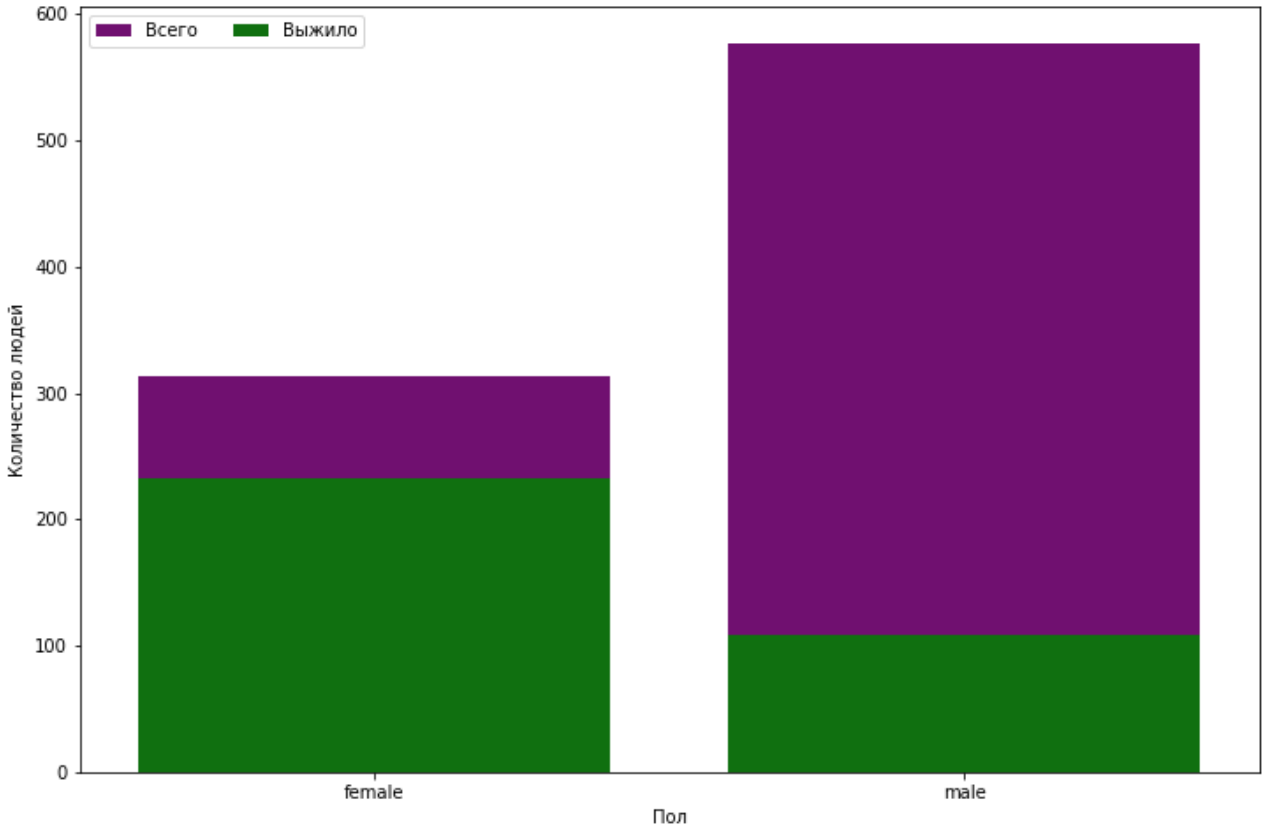

Рис.: Количество выживших/умерших мужчин и женщин

Разделение по классу билета

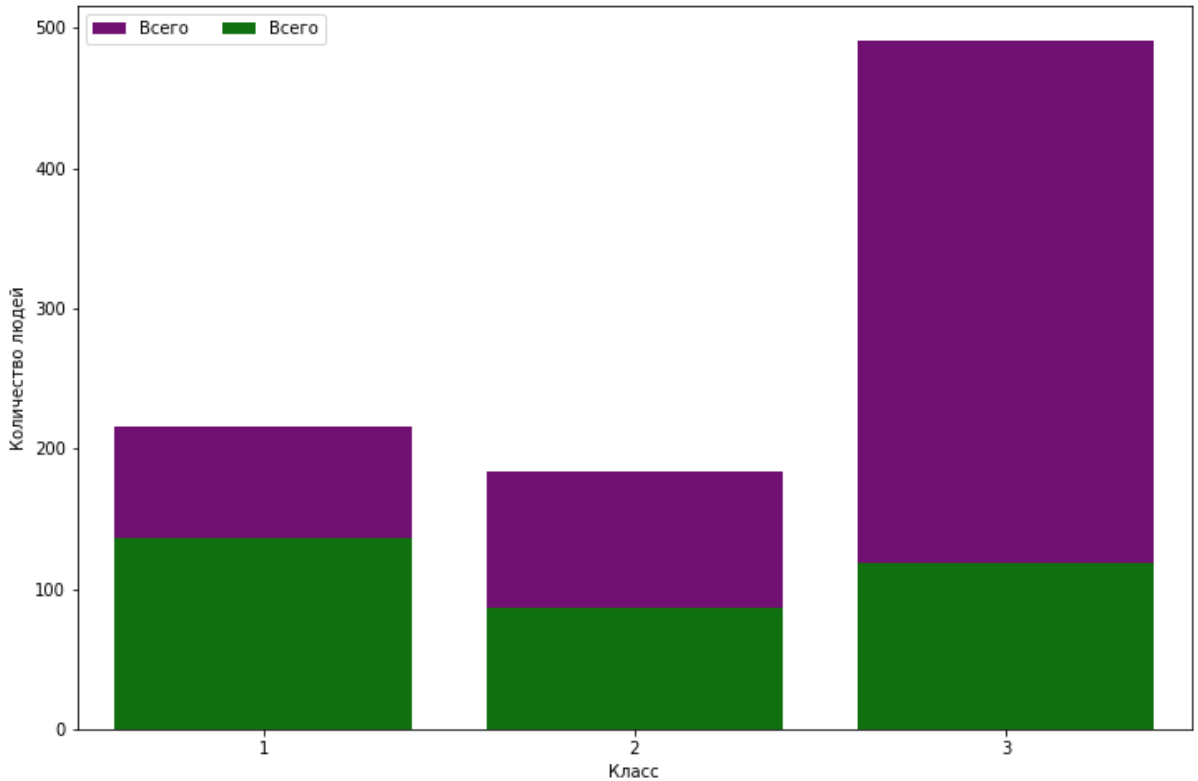

Рис.: Количество выживших/умерших в зависимости от класса билета

Корреляция признаков

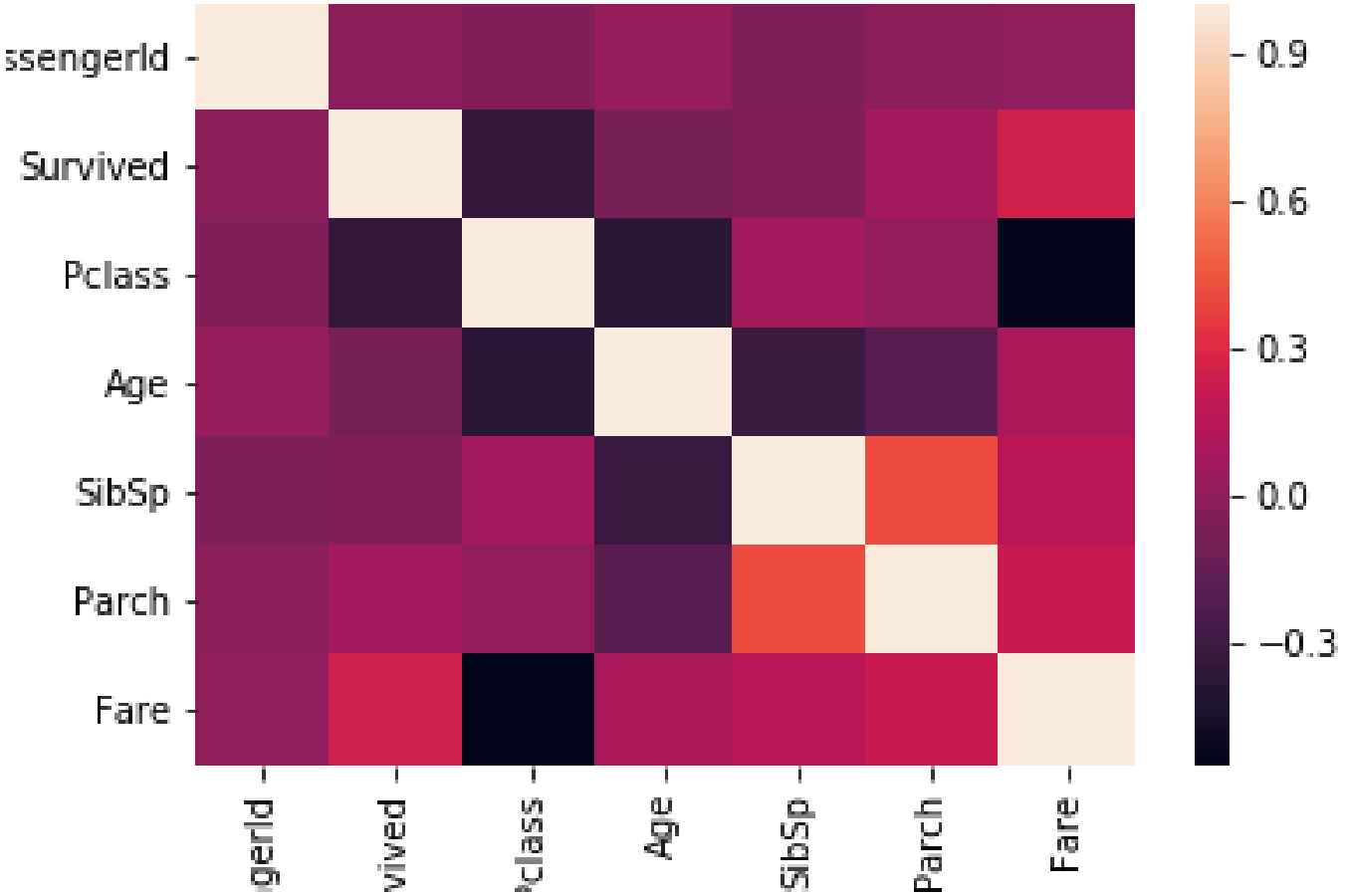

Рис.: Heatmap

Выявление новых признаков

- Присутствует ли семья человека на борту?
- Человек - ребенок?

Корреляция признаков

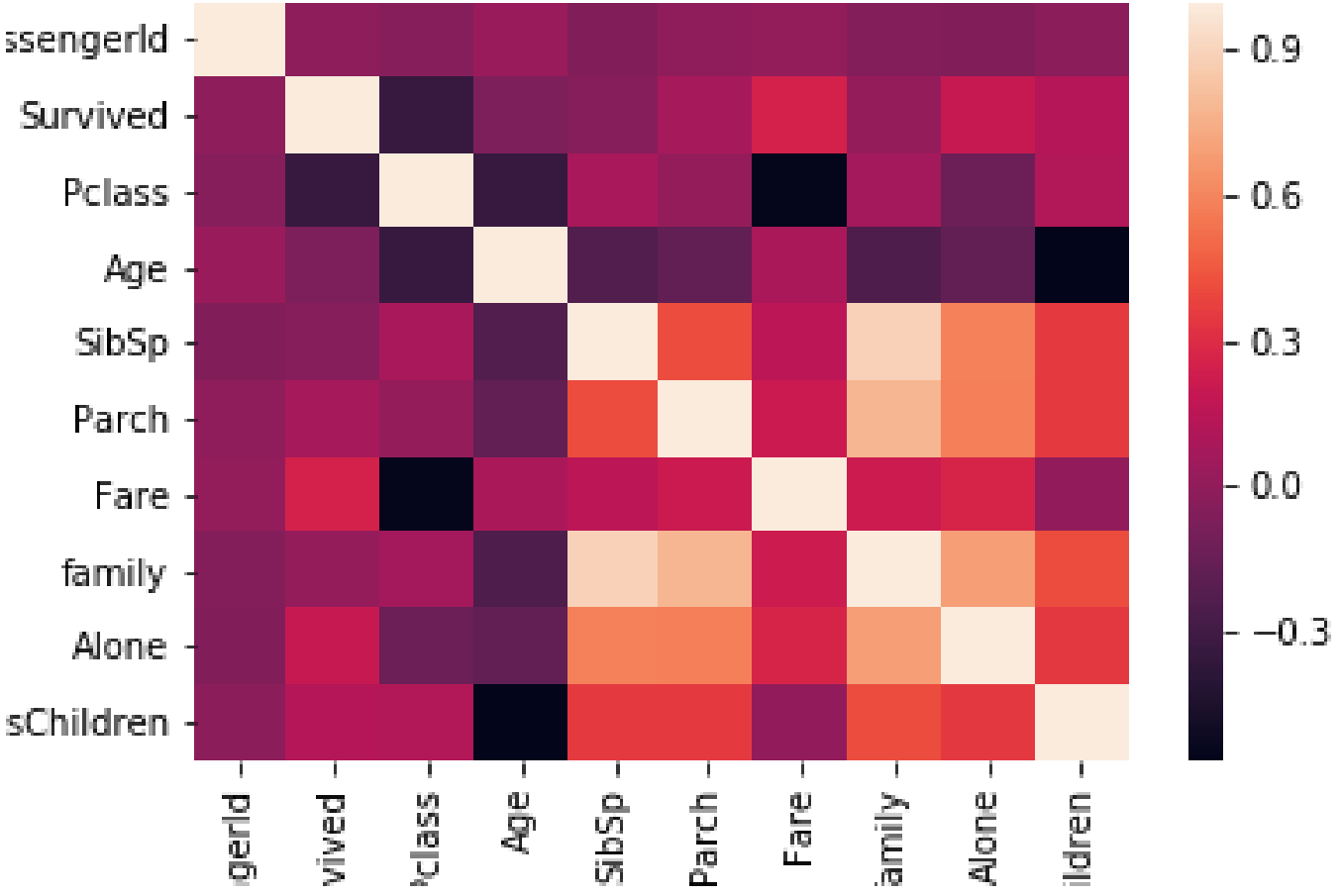

Рис.: Heatmap

Оценка моделей

- Решающее дерево: 83%
- Случайный лес: 85%
- Метод k ближайших соседей: 70%