Целевая переменная — Survived.

Наша задача — предсказать, будет она 1 или 0.

Лишние данные, не влияющие на результат:

PassengerId — судя по всему, это id, присвоенный при внесении в БД.

Поле Sex — лучше привести к значению 0,1, это немного сократит затраты.

Дополнительно хотелось бы определить влияние национальности на результат, её можно определить приблизительно по фамилии (в самом общем виде - английская и неанглийская). Для этого необходимо добавить другой столбец с категориальной переменной (например, 1 — английская, 2 — финская, 3 — славянская и т.п.). Но это отложим на будущее. Кроме того, у некоторых из пассажиров имеются титулы.

Номер помещения состоит из буквы и цифры. Буква может означать палубу или отсеки судна, значит, имеет смысл отделить её как отдельный столбец (строка), а номер каюты в отсеке — отделить в другой (количественный). Вообще-то номер — это скорее текстовый параметр, но можно предположить, что нумерация в каждом отсеке была последовательная, например, от входа, и тогда возможна зависимость целевой переменной от старшинства номера. Можно также посмотреть на «зоны», которые создаются в районе определённых номеров, возможно, что это наиболее опасные зоны на этом судне.

Ticket как представляется, не даёт полезной информации, поэтому его можно удалить. Во всяком случае для быстрого анализа он не представляет достаточно данных.

Далее по данным:

|  | PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Embarked |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | int | int | int | string | int | int | int | int | string | float | string | string |

Категориальные переменные:

Survived, Pclass, Name, Sex, Cabin, Embarked

Количественные переменные:

Age, SibSp, Parch, Fare

Выбросы

в колонке Fare на графике видны 3 выброса, где стоимость составляет 512.3292 долларов за место. Возможно, это какие-то vip-билеты, возможно, в данном случае была объединена цена за несколько билетов или один пассажир оплатил несколько кают. Тем не менее, выбрасывать эти данные не хочется, тем более, что все три пассажира, заплатившие так много, выжили. Поскольку все они плыли 1-м классом, то однозначно здесь нет счётной ошибки.

Возраст для округления приведён к целому числу.

Пропуски

Заполняем пропуски с помощью fiillna