## Kaggle竞赛 – Titanic

1. **以pd.DataFrame形式导入数据**

pd.read\_csv()

1. **数据分析特征情况EDA（exploratory data analysis）**

Datafrme.info() datafrme.describe()

善于利用matplotlib.pyplot可视化查看各特征与预测目标的关系

1. **特征工程(特征选取、缺失值填充、特征转换)**

**特征缺失解决方法：**

1. 如果缺失比例太高，可选择舍弃该特征，以免反倒带入noise
2. 如果缺失比例适中，特征属性为离散值，可选择把NaN作为一个新类别，加到特征中去

df.loc[df.fea.isnull(), ’fea’] = ‘Yes’

df.loc[df.fea.notnull(), ’fea’] = ‘No’

1. 如果缺失比例适中，特征属性为连续值，可选择给定一个步长离散化后，将NaN作为一个type加入到属性类目中
2. 如果缺失比例不多，可选择根据已有数据拟合补充

**特征值归一化/标准化：**

Sklearn.preprocessing.StandardScaler

**特征转换：**

1. 可将若干重要特征组合为作为一个新特征(FM思想)
2. **模型选取、拟合、测试**
3. 利用交叉验证测试模型效果(auc)

Sklearn. cross\_validation. train\_test\_split()

1. 先跑出一个baseline，再进行优化
2. 模型融合