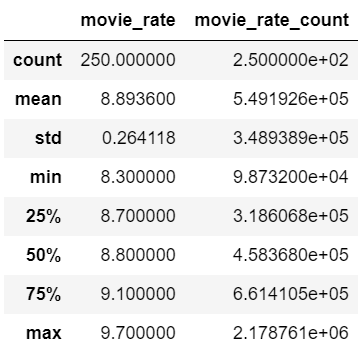
# 统计指标分析





# 可视化图表分析

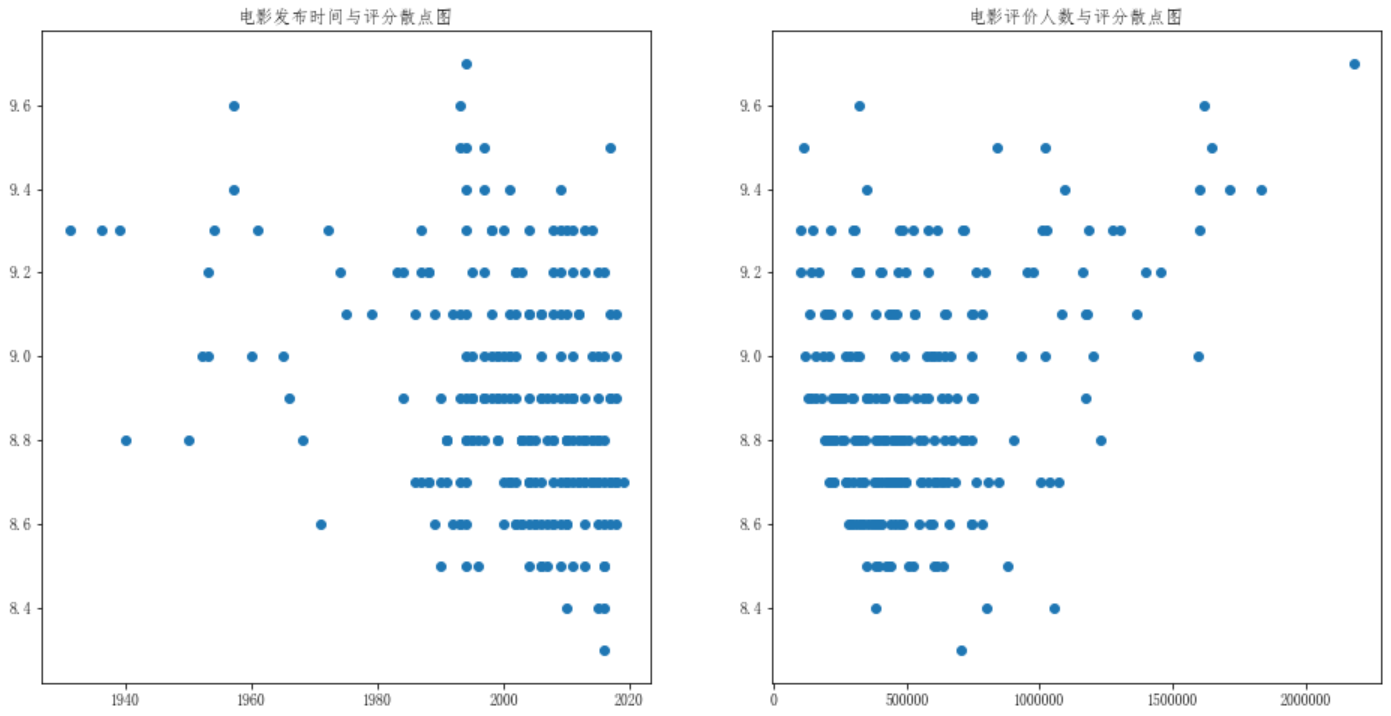
## 电影类型词云



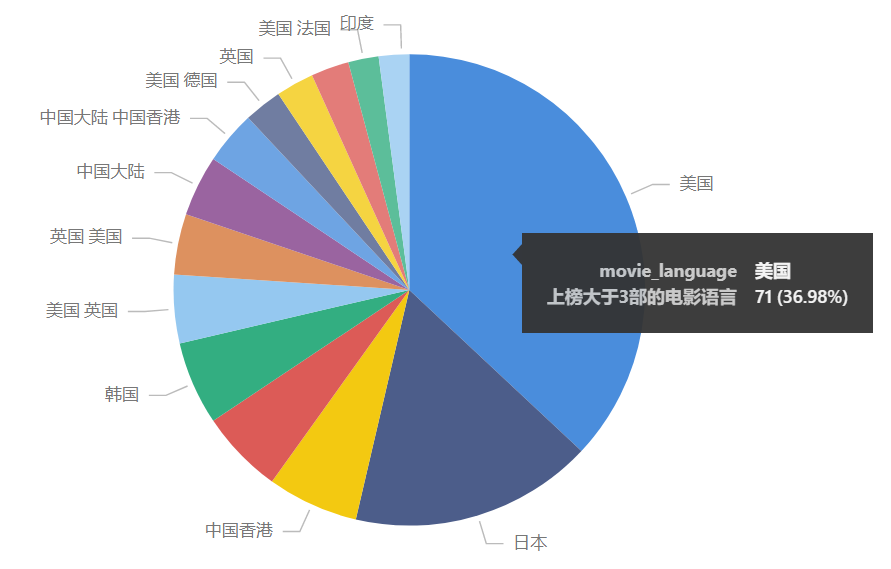
## 电影语言词云



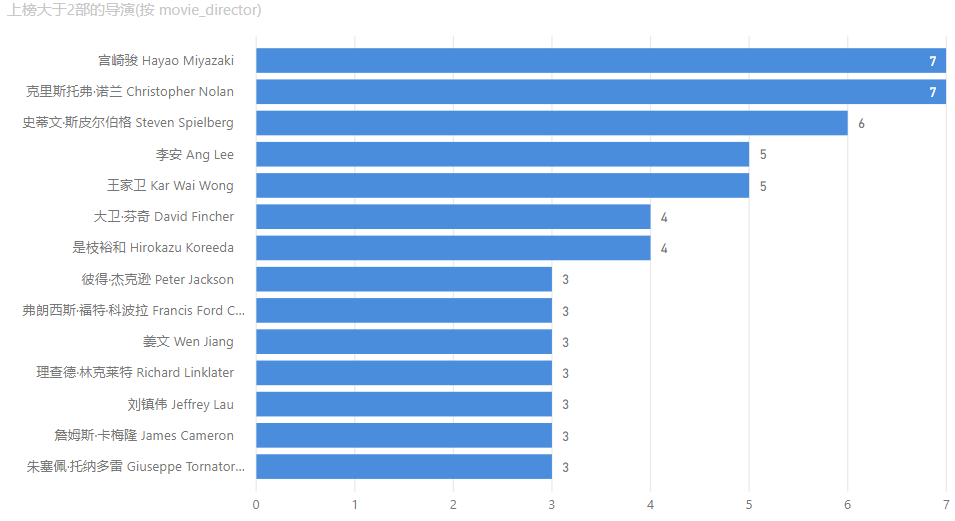
## 散点图



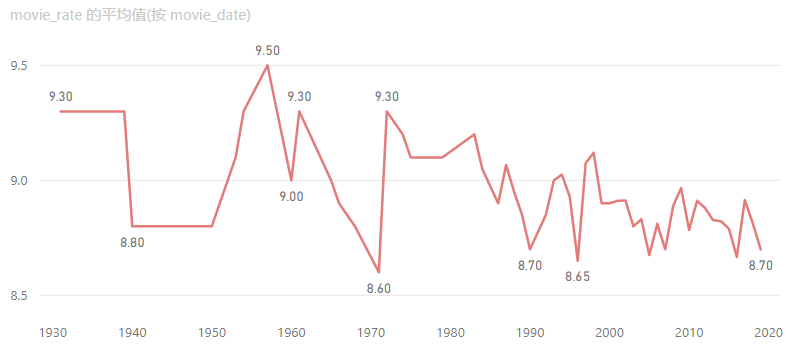
## 上榜大于3部的电影语言



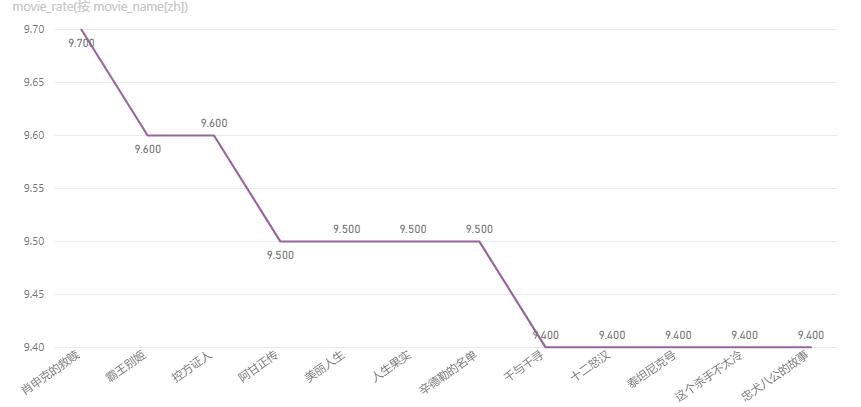
## 上榜大于2部的导演



## 电影评分平均值按年分布



## 评分大于9.3的电影



# 机器学习建模预测性分析

* **预测因变量**：电影评分（movie\_rate）
* **预测自变量**：电影评分人数（movie\_rate\_count）、电影年份（movie\_date）、电影类型（movie\_type）、电影语言（movie\_language）

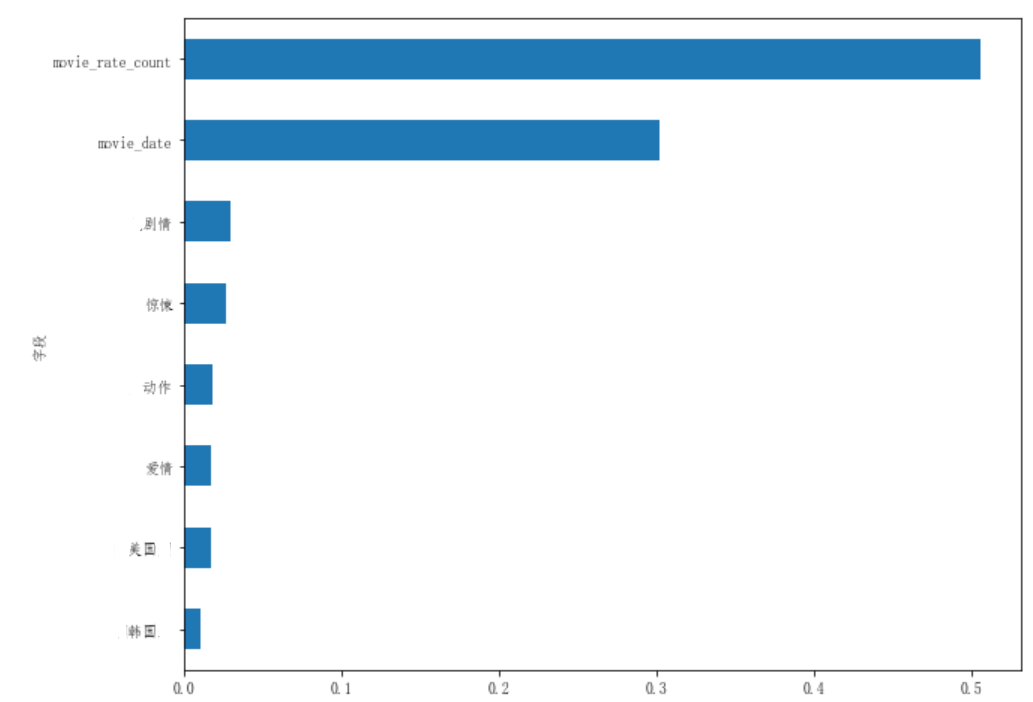
其中，由于电影类型和电影语言是离散性数据，在预处理时转化为了dummy variables

## 模型准确性分析

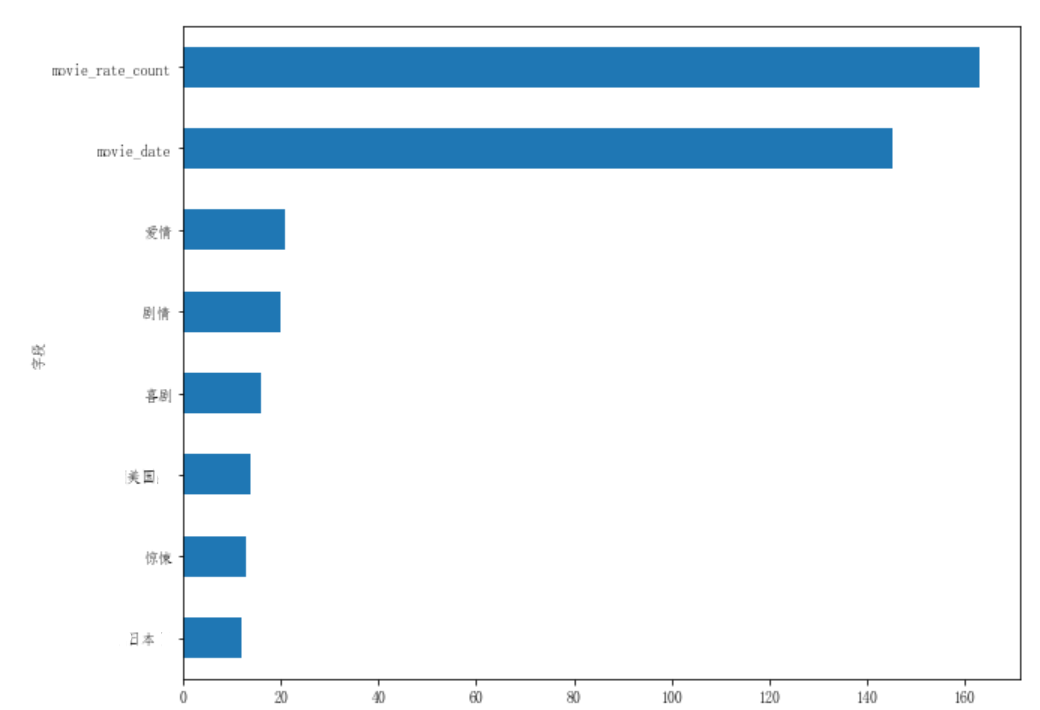
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | train | | validation | |
| ***RMSE*** | ***R2*** | ***RMSE*** | ***R2*** |
| Null Model | 0.26 | - | - | - |
| Random Forest | 0.17 | 0.60 | 0.21 | 0.29 |
| LightGBM | 0.18 | 0.53 | 0.22 | 0.24 |

## 自变量相关性排序

### 随机森林



### LightGBM 回归



## 总结

随机森林的预测准确性更好