1）introduction

* 项目背景，数据来源，audience (audience会影响未来的model选择)，项目的目的（找到某个label未来的走势等）

2）Exploratory Data Analysis and Literature review（初步数据分析）

* 先详细的说一下数据来源，最开始数据的size（比如目前是1000 x 111 + 11个label）
* 这个数据集有什么明显的问题，接下来预计用xxx方法去处理它 (缺失值特别多，只有极少feature符合正态分布，有明显的right skew或者left skew)

3）Methodology

* 首先写我们数据处理用了什么方法（knn和函数来填补空值），理由（missing value过多，直接删除会影响信息量）
* 写为什么我们选择把这些label看成是一个 一起train（选择自带multi task的model），或者为什么要分开train（比如分成11个model去做train）
* 每一步都及时更新数据集shape的变化
* Model Selection： （比如选lgr：binary的task跟lgr很契合，同时lgr比较方便把label的输出转成一个概率，对医药行业人士来说容易理解balabala）

4）Results

* 找一个基准线（最差的一个model，以后的都和它比较好在哪儿），参数需要完整的详细的写出来。还要写他的performance（f1 score）。同时要有一个错误分析，分析为什么这个模型表现得烂。可以写一个confusion matrix：针对某个label（比如false positive特别多）可能的原因是什么
* 最后给出model的最终参数，f1 score，cross validation波动大不大（看model对于train set的改变是否敏感，也就是model的generalization/variance）这块表现得好不好。）

5）Discussion

* 把模型们的performance列表格做直观对比。其次看model之间performance为什么会有些好，有些坏，和model结构有什么关联，以及和我们之前的数据处理方式有没有关系。
* 未来的改进方向：收集更多数据。（某个class的数据量特别少，多找一些那个方向的数据。
* 未来的商业建议：prediction结果也可以做可视化，通过可视化发现某个方向（某个feature和某个label）相关性很大，可以根据这个做出预测和未来方向的总结。

6）Conclusion

* 和introduction相对，用了哪些方法，得到了怎样的performance，有哪些结论，以及整体项目未来的改进方向。

7）Reference