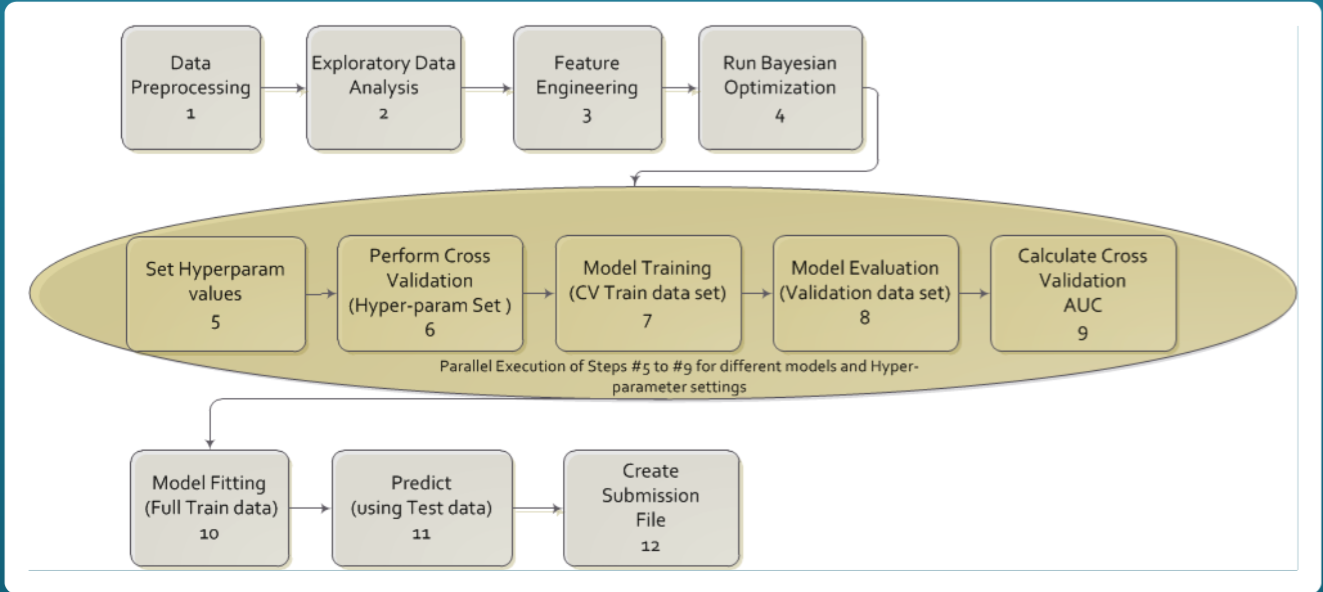
# 数据分析基本框架

## 参考：

1. <https://blog.csdn.net/han_xiaoyang/article/details/52788775>

## 数据分析一般流程——****Agile过程****



## 重点

### 特征工程

缺失值估算方法：KNN，平均值，随机数和中位数。然而，我们发现，不去补充缺失值，模型反而会有更好的表现

我们也做了PCA。结果表明，选择5个主成分只能解释总方差的66％。由于只有10个特征，PCA可能不适合这个项目。

对于每一列，我们检测基于第95百分位数的异常值，基于中位数的异常值和基于标准偏差的异常值，并用投票来决定最终结果。如果三种方法中，如果有两种方法都认为某一个点是一个异常值，那么我们就确定这个点是一个异常值。然后，我们将异常值替换为该列的中值或最不异常的异常值。至于使用哪种替换，取决于具体的列。

### 模型融合

我们为了本次挑战，构造了4个模型，并评估了他们的准确率。团队实施了一个并行流程，所有的模型同步构造。在自动优化过程中，每发现一个更优的参数集，这些参数会用于整个流程循环，即时地进行协同工作。

**简单和集成模型**

**优化的堆叠模型**

**投票分类模型**

**投票和堆叠的混合模型**

### 超参数优化

共有4中超参数优化方法：1、网格搜索 2、随机搜索 3、基于梯度的优化 4、贝叶斯优化。在这4中方法之中，我们尝试了网格搜索，随机搜索和贝叶斯优化。我们发现贝叶斯优化是最高效的，可以自动达到最优。

## 接下来的工作

1. 整理Home Credit公司背景资料
2. 逐步对特征进行探索
3. 了解贝叶斯优化以及其它优化器
4. 了解LightGBM算法
5. 了解xgboost算法
6. 了解**简单和集成、优化的堆叠、投票分类模型**