**微博互动预测方案**

题目来源：<https://tianchi.aliyun.com/competition/entrance/231574/information>

**1 Motivation**

1.1 背景

1.2 商业运用场景

**2 数据整理**

2.1 ~~剔除缺失值（无缺失值）~~、错误值等等

2.2 筛选掉噪音数据：

* 1. 全英文、符号、数字
  2. 全链接
  3. 无效内容：“点下面网址”、只有“分享图片”
  4. 字数过少的博文（如字数小于5）互动量很可能为0

2.3 在训练集中划分出validation set：2月至6月为训练集，7月为validation set

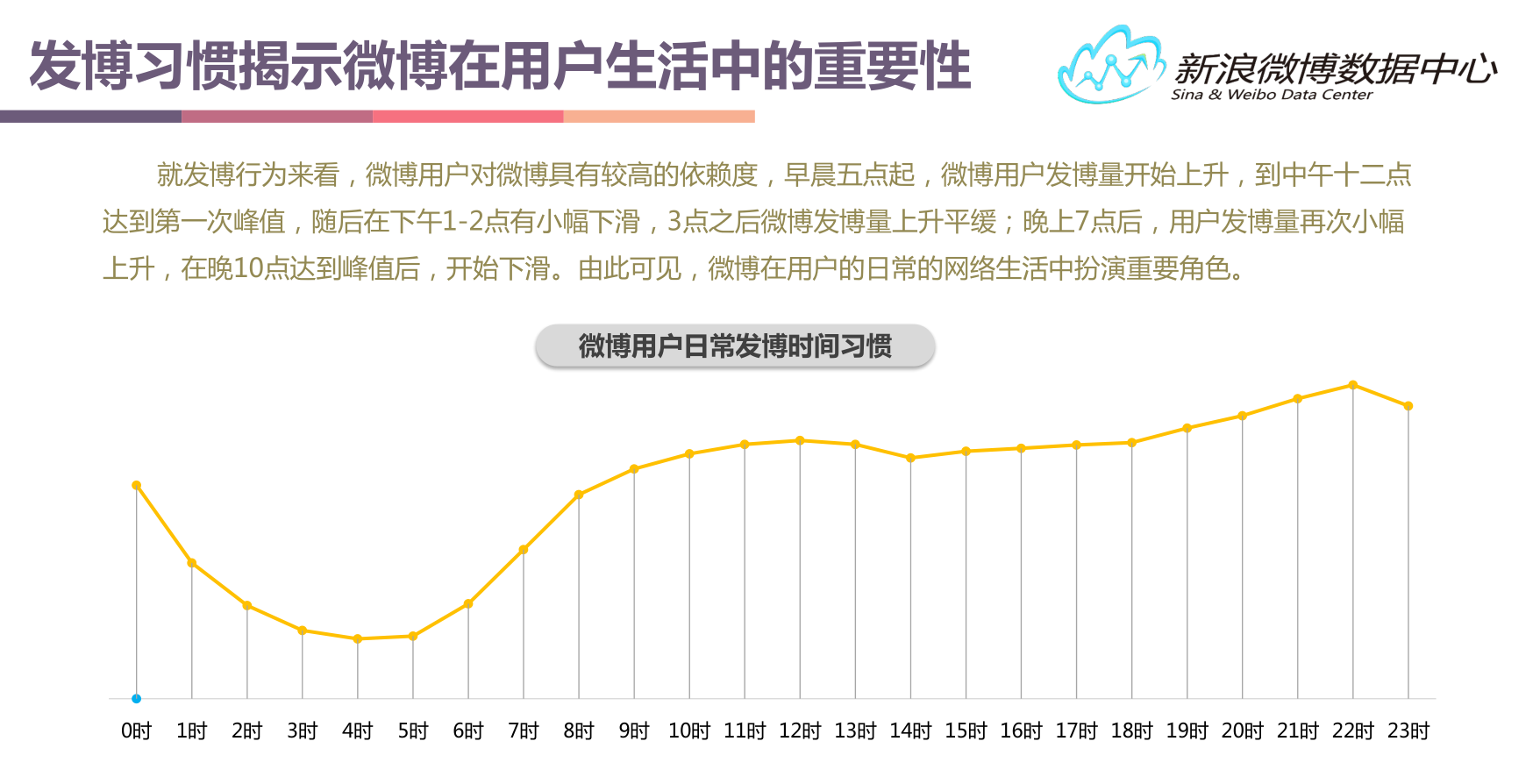
2.4 特征提取（形成一个面板数据）

1. 用户：
   1. 历史数据的 range, 平均值和方差。
   2. 在训练集中出现的次数（发博频率）
   3. 在训练集中收获转发的最高次数（估计粉丝量水平）
   4. 历史评论中位数/转发中位数（账号类型）
2. 时间（下列各项数据均为binary）：
3. 是否周中，是否周末
4. 是否节假日
5. 发文时间段（需要结合数据再最后确定）：

早上：2 am-6 am, 6 am-9 am, 9 am- 12 am

下午：12 pm-15 pm, 15 pm -19 pm

晚上：19 pm- 23 pm, 23 pm-2 am



[（2015微博用户发展报告）](https://data.weibo.com/report/reportDetail?id=297)

1. 工作日与节假日可能存在交互效应（作息不同）
2. 在训练集中出现次数 / 第一条为微博到最近一条微博时间
3. 文本特征：
   1. 微博长度（原始长度，中文长度）
   2. 是否为英文内容（是否为非中文内容）
   3. 是否为转发（没有发现样本）
   4. 是否为分享
   5. 是否为回复自动发博
   6. 是否为互动内容（“//”）
   7. 是否包含书名号（对作品的讨论）
   8. 是否 @ 他人
   9. 是否为投票
   10. 是否为抽奖
   11. 是否包含表情
   12. 是否包含视频
   13. 是否包含链接
   14. 【欧阳】是否为第三方平台互动消息（“我在#xxx”）



* 1. 【欧阳】是否包含【】/是否有标题（很可能是新闻）\* 训练主题分类器会不会太麻烦了
  2. 【欧阳】是否为娱乐内容\* 训练主题分类器会不会太麻烦了是否包含hashtag#
  3. 【欧阳】是否为广告（是否包含天猫、淘宝、购物券、折扣、优惠、红包等等词语）
  4. 【欧阳】是否包含当年的百度热词：
  5. 【欧阳】TF-IDF 关键词频率
  6. 【欧阳】是否包含样本里出现的高频词语（非广告词）
  7. ~~情绪特征？\*~~

**3 数据分析（基于特征）**

**3.1 描述性统计 (overview of data，可以想象成是微博的用户报告)**

**目的：**1）介绍数据集 2）通过简单观察，得到显著的特点（用于后续预测），例如工作日凌晨3点的微博平均转发量几乎为0；广告型的博文互动量极低等等 3）增加展示的趣味性。 展示时，只展示选取最显著的特征。

**概述：**

微博数量，时间范围（2015年2月-7月，每天，24小时），用户数量

每篇互动量的频率分布直方图

**用户维度：**

* 用户的发博数量、互动量、粉丝量的平均值，max值，min值（表格描述）
* 每月发微博次数的用户分布（0-5次，5-10次等等，饼图）
* 每篇微博的平均互动量的分布（饼图）

例如每篇微博转发量为0的用户的比例，0-5的用户比例等等

* 用户粉丝数量（最大转发量）分布比例（饼图）

**时间维度（类似于气温展示）：**

微博数量：

* 一周内每天的微博数量（柱状图）
* 一天内24小时的微博数量（柱状图）

工作日24小时的微博数量

周末24小时的微博数量

互动量（平均值，最大值，最小值）：

* 一周内每天的平均互动量（柱状图）
* 一天内24小时的平均互动量（柱状图）

工作日24小时的平均互动量

周末24小时的平均互动量

**文本维度：**

每个文本特征的分布，及其与互动量的关系

例如：微博文本长度的分布（0-5字的比例、5-15字的比例等等，表格展示），不同长度的文本的平均互动量分布；广告性的微博可能互动量极低。

**3.2 算法比较**

零值预测：作为baseline

中位值预测：老用户用中位数，新用户用0

普通回归预测

决策树回归树

随机森林回归树

综合模型：

1. 筛选一些显著互动量为零的数据：不活跃用户、新用户，非中文（可以），字数过少（小于等于5个字）（可以），自动发布的微博（支持），分享微博，互动内容（可以），投票（可以），抽奖，广告（可以）。
2. 用 logit 回归判断互动量为零的数据（尽可能降低TN比率），使用ROC曲线判定阈值
3. 对于非零数据，使用随机森林算法，用CART的回归树判定互动量

**4 结果展示**

**4.1 线上竞赛得分**

**4.2 线下测试集得分/准确率**

**5 分工计划**

特征提取（7号晚上交代码）：

用户特征：安蕾， 时间特征：妍戎，文本特征a-m:永乔，文本特征n-t: 欧阳

**6 疑问**

6.1 转发量、评论量和点赞量的关系如何？是否要做三次模型测试？

6.2 样本可能面临严重的非均衡分布问题，需要重采样？

一般认为，零互动量的样本数量：非零互动量的样本数量>4，就算严重