# Machine Learning Engineer Nanodegree

## Capstone Proposal

## 毕业项目-开题报告 UDA学员 李欣

**2019.2.13 农历（己亥年）猪年 北京**

### Domain Background

这是一个来自于Kaggle的真实竞赛项目，项目要求预测Rossmann的日常销售。Rossmann是一个在7个欧洲国家经营着3,000多家药店。目前，Rossmann商店经理的任务是预先提前六周预测他们的日常销售。商店销售受许多因素的影响，包括促销，竞争，学校和州假日，季节性和地方性。成千上万的个体经理根据他们独特的情况预测销售情况，结果的准确性可能会有很大差异。

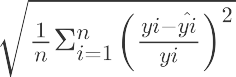
在Kaggle的比赛中，Rossmann向您提出挑战，要求您预测德国各地的1,115家商店每周销售6周。可靠的销售预测可以使商店经理能够创建有效的员工时间表，从而提高生产力和动力。通过帮助Rossmann创建一个强大的预测模型，您将帮助商店经理专注于对他们最重要的事情：他们的客户和他们的团队！

之所以选择这个项目，是因为该项目与我的实际工作中亟待解决的问题同出一辙。

### Problem Statement

该项目的具体任务是在1,115家Rossmann商店的历史销售数据中预测测试集的“Sales”列，这属于线性回归类的需求。可以通过构建一个有监督学习类的模型来减少预测值（ŷ）与实际值（y）之间的误差来解决本需求。

使用Kaggle的RMSPE 函数来验证真实的销售数据与预测数据的差异性。RMSPE 函数：



**Datasets and Inputs**

根据Kaggle提供的数据，现有1,115家Rossmann商店的历史销售数据（如下表，训练集：train.csv；测试集:test.csv；商店的特征集:store.csv）。 测试集的“Sales”列就是实际值（y）。 数据集中的某些商店暂时关闭以进行翻新。由于指标中yi不能为零，因此在数据清理过程中将排除实际销售额为零的数据。

|  |  |  |
| --- | --- | --- |
| 数据集 | 内 容 | 数 量 |
| train.csv | 销售⽇、销售额、顾客数、促销、营业、国家假期、学校假期 | 1017209 |
| test.csv | 销售⽇、顾客数、促销、营业、国家假期、学校假期 | 41088 |
| store.csv | 商店的特征信息，包含：商店类型，等级，竞争者信息，促销活动 | 1115 |

|  |  |
| --- | --- |
| 特征名 | 含义 |
| Store | 商店编号，1-1115 |
| Sales | 日销售额 |
| Customers | 日顾客数 |
| Open | 是否营业，0关门/1开门 |
| StateHoliday | 州假日，a=全部假日/b=复活节/c=圣诞节 |
| SchoolHoliday | 学校假日是否开门，0关门/1开门 |
| StoreType | 商店类型 ，a,b,c,d |
| Assortment | 商店评级 ，a,b,c |
| CompetitionDistance | 与最近的竞争者之间的距离 |
| CompetitionOpenSinceYear | 最近的竞争者开业的年份 |
| CompetitionOpenSinceMonth | 最近的竞争者开业的月份 |
| Promo | 当天是否促销，0否/1是 |
| Promo2 | 商店是否参与长期促销，0否/1是 |
| Promo2SinceYear | 商店参与长期促销开始的年份 |
| Promo2SinceWeek | 商店参与长期促销开始的日历周 |
| PromoInterval | 长期促销的月份，Feb/May/Aug/Nov |

### Solution Statement

解决方案分为:1、数据分析；2、模型搭建；3、模型评估及调参三个阶段。  
 1、数据分析和清理

·为了获取数据中的有效特征，对数据进行可视化分析，获得数据分布特征，发现数据之间的内在联系并排序有效特征。

·数据预处理，独热编码、寻找并增减异常数据、补全缺失值、规范特征格式、修缮特征字段等。

2、模型搭建

采用XGBOOST建模，预测出 1115 家门店的未来销售额。

1. 模型评估及调参

### 使用GridSearchCV 方法或者贝叶斯优化来进行模型的参数调优，寻找最佳参数组合，以使模型绩效最大化。

### Benchmark Model & Evaluation Metrics

首先，搭建一个实现项目基本需求的基础模型，为之后的模型优化提供基准。

其次，使用LightGBM、Xgboost等算法训练数据，再依据Kaggle推荐的衡量方式：RMSPE来评估模型，最后选出最佳模型再对测试集的商店销售额进行预测。

最后，将预测成绩提交到Kaggle上参加竞赛，至少达到 leaderboard private 的 top 10%以内。实际对于测试集的Rmpse ，需要至少达到 0.11773分。

### Project Design

**第一阶段，数据分析与预处理**

1：对数据进行独热编码、寻找并增减异常数据、补全缺失值、规范特征格式、修缮特征字段等。

2：利用pandas、matplotlib 和 seaborn对数据进行探索性的可视化分析，深刻理解数据的内在联系。

3：对数据进行一些统计上的梳理，获取数据中的分布特征，并排序有效特征。

4：特征选定，将store与train和test数据集合并，并Drop掉冗余和无效特征。

**第二阶段，模型搭建**

1：选用一些常见的模型（LightGBM、Xgboost）进行训练。先使用模型默认的参数进行训练尝试，如多项式的线性回归和逻辑回归，在第三阶段经过模型评估之后，再决定使用更适合的模型。

2：Model Training，确定好自己训练和测试集，已经分好了，所以此步骤就相当于 Cross Validation，知道自己model有没有overfit。\*

3：Ensemble Generation，尝试使Bagging来用训练数据的不同随机子集来训练每个 Base Model，最后进行每个 Base Model 权重相同的 Vote。也可以尝试使用Boosting来根据上一个迭代中预测错误的情况修改训练样本的权重。或者用Blending将不相交的数据训练不同的 Base Model，将它们的输出取（加权）平均，或者使用较为复杂的Stacking。最终如何还是要看自己选择的模型和数据的具体情况了。

**第二阶段，模型评估及优化**

1：使用GridSearchCV 方法或者贝叶斯优化来进行模型的参数调优，寻找最佳参数组合，以使模型绩效最大化。

2：通过控制模型的复杂度、加强模型的鲁棒性来控制模型过拟合。

**第四阶段，结论与展望**

1：使用最后优胜的模型生成预测数据，提交到kaggle上，完成Kaggle项目竞赛。

2：对项目进行总结，分析特征工程、模型选择与调参阶段的理解。

3：对项目进行展望，分析不足与期望。