# **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多家日用品的连锁企业。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。

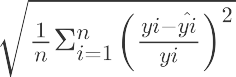
在Kaggle的比赛中，Rossmann向您提出挑战，要求您预测德国各地的1,115家商店每周销售6周。可靠的销售预测可以使商店经理能够创建有效的员工时间表，从而提高生产力和动力。通过帮助Rossmann创建一个强大的预测模型，您将帮助商店经理专注于对他们最重要的事情：他们的客户和他们的团队！

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。

训练集train.csv：2013.01.01-2015.07.31，共942天;

测试集test. csv：2015.08.01-2015.09.17,共48天;

商店信息数据集store.csv: 1115条数据，共1115家商店的信息。

|  |  |  |
| --- | --- | --- |
| 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**

技术方案分为；1、数据分析；2、模型选择；3、算法评估及调参三个阶段。  
 1、数据分析和清理

·特征变量的分布

·目标变量的分布

2、模型选择

也是最有意思，最困难的地方，需要了解这个问题的背景和各个特征的意义，根据各个特征的实际情况和常识来补充缺失值，然后找出那些有密切实际联系的特征，将其转换为新的更直观更适合算法模型的特征。  
 模型选择方面，该项目是一个监督学习的回归问题，主流的监督学习回归算法都是可以适用的，通过设定一个基准分数尝试多个回归模型来比较选择其中表现最好的。模型训练时间在建模的时候成本很高，但是在建模成功使用其进行预测的时候建模成本并不影响。  
 模型的评估和优化是一个永恒的主题，训练建模的时间制约了尝试训练的次数，可以采用cv的方式寻找到最佳的参数。但是最好的优化还是在第一步数据整理上，一个项目的数据整理决定了模型的上限，再好的模型和优化都只是无限逼近这个上限。  
 目前我的XGBOOST建立的模型达到了我的预期，在通用场景的话可以按照这个思路来训练建立模型，但还是要考虑每个场景的具体情况。

对于该项目我觉得还有些特征的理解我有疑问，商店里的商品那么多做Promo2的商品是哪个牌子的哪个型号，难道预测的是笼统销售活动；竞争对手开店时间1900应该是个默认值所以应该不是很准确；顾客数在预测工作中应该是很重要的，而我在这个模型里却没有想到改如何使用。

### **Benchmark Model**

(approximately 1-2 paragraphs)

In this section, provide the details for a benchmark model or result that relates to the domain, problem statement, and intended solution. Ideally, the benchmark model or result contextualizes existing methods or known information in the domain and problem given, which could then be objectively compared to the solution. Describe how the benchmark model or result is measurable (can be measured by some metric and clearly observed) with thorough detail.

### **Evaluation Metrics**

(approx. 1-2 paragraphs)

In this section, propose at least one evaluation metric that can be used to quantify the performance of both the benchmark model and the solution model. The evaluation metric(s) you propose should be appropriate given the context of the data, the problem statement, and the intended solution. Describe how the evaluation metric(s) are derived and provide an example of their mathematical representations (if applicable). Complex evaluation metrics should be clearly defined and quantifiable (can be expressed in mathematical or logical terms).

### **Project Design**

(approx. 1 page)

In this final section, summarize a theoretical workflow for approaching a solution given the problem. Provide thorough discussion for what strategies you may consider employing, what analysis of the data might be required before being used, or which algorithms will be considered for your implementation. The workflow and discussion that you provide should align with the qualities of the previous sections. Additionally, you are encouraged to include small visualizations, pseudocode, or diagrams to aid in describing the project design, but it is not required. The discussion should clearly outline your intended workflow of the capstone project.