**总体设计方案**

**编制:雷邓君**

**项目名称:公平价二手车估值升级**

**本文件修订历史信息:**

|  |  |  |  |
| --- | --- | --- | --- |
| **更新日期** | **修订者** | **批准人** | **主要修订摘要** |
| 2018.10.31 | 雷邓君 | 万可文 | 新编写文件 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

# 1.总体目标

## 1.1总体目标

估值系统:能通过品牌,车系,款型,公里数,城市,上牌时间特征预测该车当前的车商收购价,车商真实售价,车商建议标价,个人交易价,最低成交价,认证二手车价,4S店置换价,车商批发价,平均买车价,平均卖车价十类价格.每种价格分四个车况,[优秀,良好,一般,较差].

残值系统:能通过品牌,车系,款型,公里数,城市,上牌时间特征预测该车当前的车商收购价,车商真实售价,个人交易价未来12个月每个月的残值和未来三年每年的残值,以及过去6个月这三类价格的历史走势.

# 2.需求分析及功能设计

## 2.1子系统划分

估值大系统分成三部分:爬虫系统,款型匹配系统和估值算法系统。

##### 2.1.1 爬虫系统

功能介绍:抓取两部分内容,一是汽车之家的款型库作为我方标准库的补充和新款车4S店的实际售车价,二是抓取瓜子,人人车,优信三平台的车源数据作为估值算法系统的训练数据.详细设计请参见吴辛成编写的《爬虫系统设计》

##### 2.1.2 款型匹配系统

功能介绍:通过车源信息的文字标题,预测其品牌,车系,款型与我方标准库的对应关系.

例如文字标题:奥迪A8L 2013款 A8L 40 hybrspecid(进口)

预测品牌:奥迪

预测车系:奥迪A8L

预测款型:2013款 A8L 40 hybrspecid(进口)

##### 2.1.3 估值算法系统

功能介绍:将爬虫抓取的车源数据作为训练数据进行训练,并根据款型,上牌年份,公里数,城市预测二手车的十类价格.

## 2.2系统流程分析

1. 爬虫抓取汽车之家款型库
2. 将我方款型库与汽车之家款型库进行匹配并最终生成我方标准库
3. 将我方标准库的款型文字描述作为训练数据，使用深度神经循环网络lstm算法进行训练并最终生成品牌预测,车系预测,款型预测三大模型.
4. 爬虫抓取瓜子,人人车,优信三平台车源数据。
5. 用3生成的模型预测4抓取的车源数据与我方标准库的对应关系
6. 将5匹配上的数据进行数据探索,特征工程,生成训练数据并使用xgboost算法最终训练生成估值模型.

# 3.系统设计

## 3.1 款型匹配系统设计

##### 3.1.1 标准库设计

我方标准库设计原则是主力洋车型库,汽车之家车型库为补充.

步骤1--品牌匹配:首先程序匹配,要求品牌名称绝对一致,无法匹配的进行人工匹配,确保绝对准确.

步骤2--车系匹配:在同品牌条件下,首先程序匹配,要求同品牌下车系名称绝对一致,无法匹配的进行人工匹配,确保绝对准备.

步骤3--款型匹配:在同品牌,车系,年款,排量条件下,首先程序匹配,要求该条件下款型名称绝对一致或者最大包含,并且新车指导价要相等或者非常接近.使用算法(2范数)进行筛选

公式: min（（文字包含率-1）\*\*2 + (指导价差异比率-1) \*\*2)

无法匹配的进行人工匹配,确保准确。

##### 3.1.2 款型匹配系统设计

款型匹配系统主要针对抓取的二手车交易平台(瓜子,人人车,优信等)的车源车型与我方标准库进行匹配.通常二手车交易平台车源文字描述只包含 车系名称+款型名称,例如:奥迪A8L 2013款 A8L 40 hybrspecid(进口).款型匹配系统设计的目标就是通过该文字序列描述预测其品牌,车系,款型。这种类型的预测是典型的序列预测,目前针对该类预测问题最成熟和常用算法就是lstm(基于tensorflow框架).

步骤1:因这类序列文字描述不存在词性分析,lstm的应用相对简单,可以直接使用单个汉字作为词向量.根据之前生成的标准库,将品牌,车系,款型的文字描述生成语料库词向量word2vec和映射器。如此操作的原因是机器学习算法只认识数字,将 奥--[010101,111111,1010101] 迪--[010101,111111,1010111] 等中文字符转换成算法认识的矩阵向量。

步骤2:将标准库的车系和款型文字描述组装成lstm训练的序列训练数据,分别生成品牌预测,车系预测,款型预测三类模型.当前模型分成三层, lstm层, dropout层, dense层。

步骤3:使用2生成的模型预测二手车交易平台车源数据的品牌,车系,款型与我方标准库的对应关系.并生成标准车源数据供估值模型算法使用.

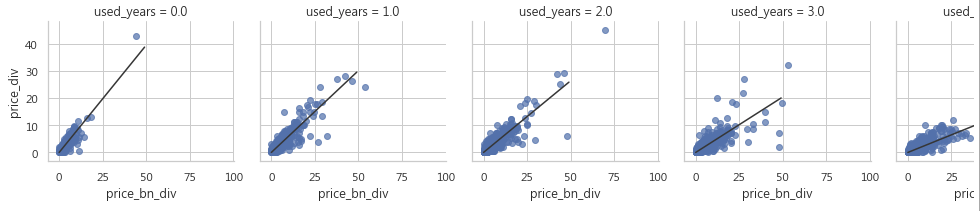
## 3.2估值系统设计

估值系统的目标是预测二手车的十类价格，机器学习内有个不成文的规定或者说是一个成熟的经验.如果训练数据空间能够覆盖超过95%，直接使用xgboos或者深度学习算法效果较好。如果训练数据量达不到该标准,那么最好降维或者降低特征类别.例如 我方标准库款型37000个,城市352个,平均车龄10年,10类价格,那么达到数据覆盖的要求就是 37000\*352\*10\*10\*0.95 = 12.4亿的数据量,但实际上本次系统升级有效数据只有98万,不到1%,远远低于经验值.直接使用深度学习算法或者xgboost效果是很差的.

所以本系统依据xgboost的数学原理和之前的市场调研,重新复现算法逻辑.人工进行树叉分支选择,每一个特征都单独分析.如此操作的前提是非常熟悉市场逻辑和数据可视化的确证.以下特征分析非常具有代表性,可以理解为树叉分支的选择.

3.2.1 指导价差拟合曲线

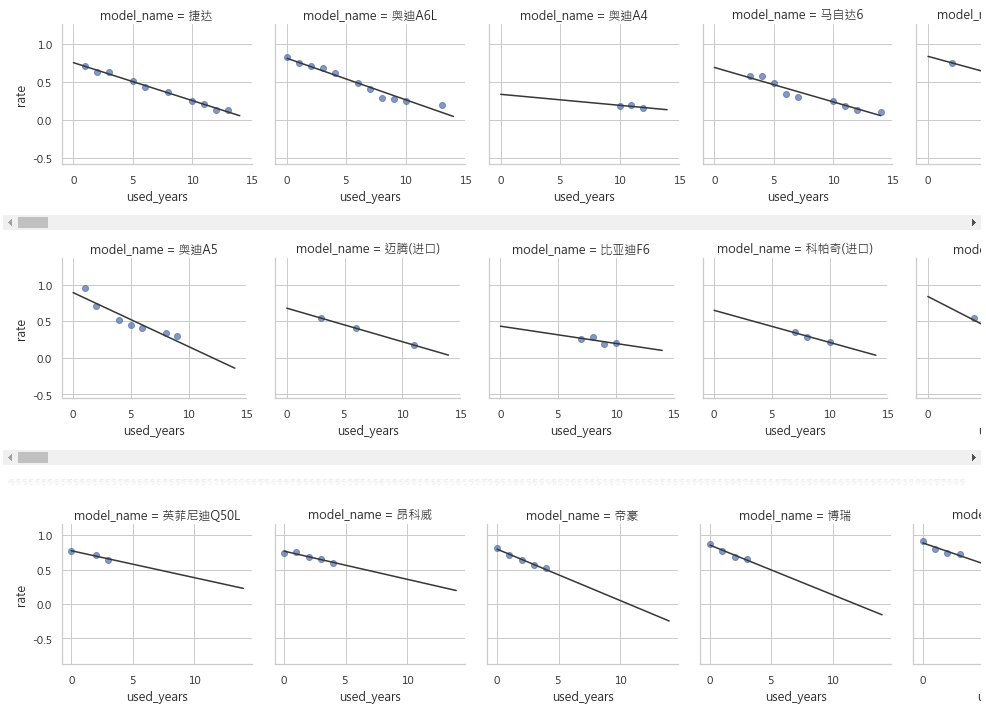
根据图形可视化,同车系同年款的车源,指导价差对于价格预测的权重随车龄呈衰减趋势.也就是说高配与低配车,随着车龄的增长,价格差异越小.训练数据展示如下:



从图像上可以看出,同车龄下指导价差与价格差呈现明显的一次曲线,当前版本也是通过一次曲线进行最终的拟合

3.2.2 车系保值率拟合曲线

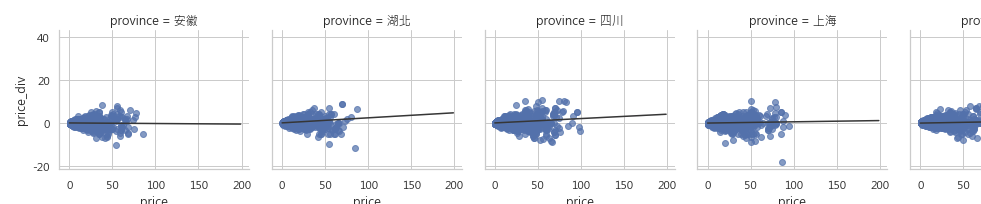
根据图形可视化,几乎所有的车系的低配车的保值率都呈现如下衰减形式:



从图上可以看出,几乎所有车系不同年款的低配车的保值率在0-6年呈现高衰减的指数曲线,6年以上的呈现低衰减的指数曲线，当前版本通过分段函数进行拟合.

3.2.3 省份差异拟合曲线

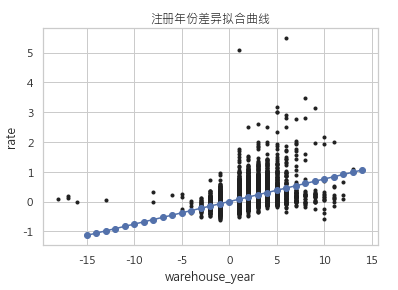
因同省份内城市差异较小,将城市分类提升为省份分类,省份差异的拟合,目前拟合的差异不是很明显,整体表现为一线城市豪华车市场占有率高,保值率偏高;二线以下城市豪华车占有率低,保值率低.中端车,低端车分析类似.因目前没有抓取到相关的其他数据,暂时以此作为分析前提,图像如下:



当前系统根据豪华车,中端车,低端车分类进行省份差异的拟合。

3.2.4 注册年份差异拟合曲线

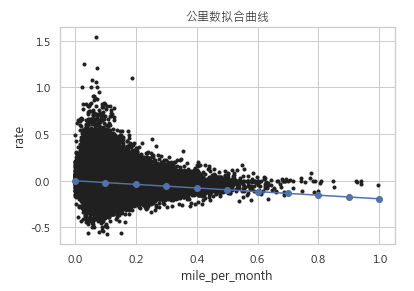
在中国同一年款车，不同年份上牌价格是有差异的,差异对最终价格的权重也是有一定的影响,图像如下:



当前系统最终拟合为一次曲线.

3.2.5 公里数拟合曲线

非营运车公里数在没有超过一定值时，对最终价格的影响权重其实很小,图像如下:



当前系统还是以一次曲线作为拟合曲线.

3.2.6 十类价格差异

因目前抓取的数据只有车商零售价，没有其他价格的数据,无法进行数据分析.目前只是通过人工设定的价格比率进行四则运算。

3.2.7 最终预测价格

低配车全国均价+指导价差拟合+省份差异拟合+注册年份差异拟合+公里数拟合+十类价格人工比率 = 最终十类价格

# 4.关键算法介绍

## 4.1 word2vec介绍

请参见网址:https://www.jianshu.com/p/471d9bfbd72f

## 4.2深度神经循环网络lstm介绍

请参见网址:https://blog.csdn.net/yoyofu007/article/details/80361422

## 4.3 xgboost算法介绍

请参见网址:http://blog.itpub.net/31542119/viewspace-2199549/

# 5.技术选型

语言:python3 使用原因:简练,易于阅读安装扩展,交互性,多范式,多平台,多框架支持

框架:tensorflow 使用原因:TensorFlow是由Google公司开发的用于实现机器学习的一种架构,是大公司背景,文档完善,使用人多.介绍文档:https://blog.csdn.net/qq\_34580082/article/details/79633535

部署:微服务+ubuntu+docker 使用原因:服务简单,只关注一个业务功能,每个微服务可由不同团队开发,微服务是松散耦合的,可用不同的变成语言与工具开发.介绍文档:https://blog.csdn.net/wuxiaobingandbob/article/details/78642020?locationNum=1&fps=1

# 6.应用安装和验收

## 6.1 安装

请参见《估值模块使用说明》

## 6.2 验收

交付python估值调用包