### 中储智运杯智慧物流算法大赛-定价竞赛

**——YGSSJ小组**

**组成员：姚栋（组长）、史志通、耿闯、盛文杰、计芳斌**

### 应用场景分析

**公司背景介绍：**中储南京智慧物流科技有限公司，组建精英开发团队倾力打造中国第一家直营式物流电商平台-"中储智慧运输物流电子商务平台"(简称"中储智运")

**产品介绍：**通过智慧物流分析与预测技术，帮助货主制定不同的发货决策，帮助承运司机获得线路热门货物的货量预报，提前做好承运准备。

**产品理念：**平台通过返程车辆的有效利用，最大化实现合理运输，减少中间环节，为广大货主提供一个价格透明、快速安全的寻车途径。

**平台模式及会员制介绍：**

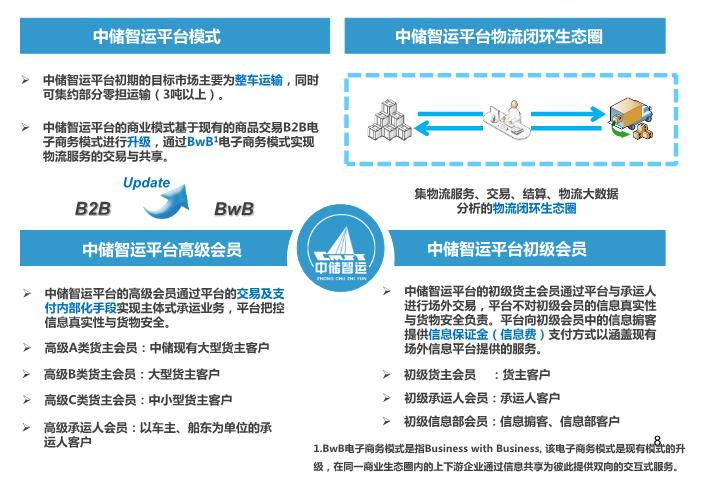


图1 平台模式及会员制介绍

**运营流程图：**

图2 运营流程图

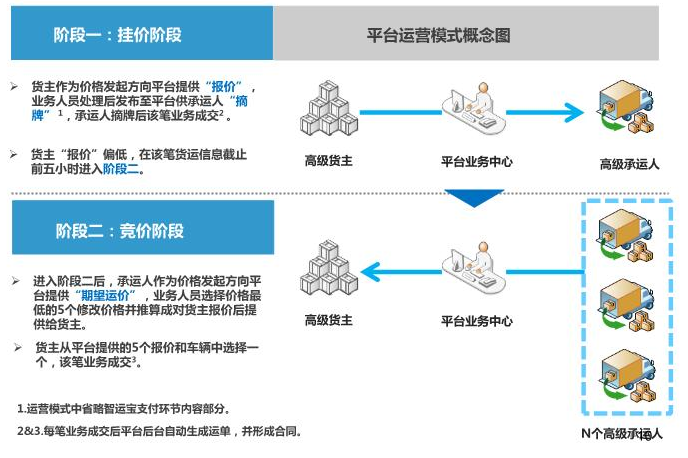
**报价模式：**

图3 报价模式图

### 数据处理与建模框架

基于应用场景分析的结果，小组讨论后决定分三步进行处理，框架图如图2所示。

首先进行数据的采集，根据经纬度信息补充油价，地区，天气的数据。采集油价的策略是填充对应地区的省市直辖市 的油价。采集天气信息的策略为首先对区域所属县级市、县或者区的天气进行采集，如果采集不到则上升一级行政区进行采集，直到省级行政区。

第二步进行数据的预处理和特征工程，数据预处理主要是对类别数据进行哑变量变换；连续数据进行z-score归一化；缺失值处理；运价的异常值处理。特征抽取则是将预处理的好的数据进行xgboost建模，对相关属性进行进行特征筛选，最终确定了49个特征。

第三步则是对确定好的数据进行建模，经过反复考虑确定使用神经网络建模。本小组最终确定使用三层神经网络的模型。



图4 框架图

### 数据采集

本小组数据采集主要是使用urllib进行网页抓取以及使用BeautifuSoup进行网页解析。主要思路为，根据提供的经纬度数据使用高德API获取相应详细地址，获取地址后再将地址作为输入获取相应地址的相应日期的天气和油价数据。

其中天气数据采集的种子页为<http://lishi.tianqi.com/>，首先根据地址进入相应地址的历史天气网页，再根据提供的日期数据去相应日期网页进行天气数据网页的下载和解析，以获得结构化天气数据。

油价数据采集的种子页为<https://energy.cngold.org//list_427_all.html>，类似的也是根据地址和日期去下载相应的网页然后进行解析获取结构化油价数据。

### 数据预处理与特征抽取

基于对数据应用场景和比赛案例背景的分析，对训练数据以及测试数据均进行以下的预处理，以便得出模型所需要的特征。

1.对于日期数据，本小组设想了两种思路：一是考虑到四季不同对运价可能造成影响，于是将一年的日期数据，按1-3月，4-6月，7-9月，10-12月分成四个季节。

二是考虑到节假日对于运价可能造成影响，于是将一年的日期数据区分成节假日和非节假日。

2.订单类型和交易类型经过本小组对业务背景的分析，认为均会对运价产生影响，且订单类型和交易类型都是二分类的变量，因此未做直接处理。

3.对于一级品类和详细品类这两个变量，在业务分析和数据分析的基础上，本小组发现仅使用一级品类能够得出更好的结果，并且一级品类是一个多分类变量，因此将其处理成哑变量，作为特征。

4.对于车型数据，考虑到不同车型的载重，长宽高，价值和折旧等都有不同，因此这些因素也会对运价产生影响，并且车型也是一个多分类离散变量，因此将其处理成哑变量，作为特性。

5.对于运距，车长和货值这三个连续变量，考虑到这些因素都会影响到最后的的运价，尤其是运距，和运价呈显著的正相关关系，因此对这三个连续变量进行归一化处理，作为特征使用。

此外，本小组根据原始数据提供的经纬度，使用网络爬虫，爬取了每个经纬度所代表的城市区域，并且还爬取了各个地区的天气和油价信息，以下是对这几个数据的预处理过程。

6.对于爬取得到的地理区域数据，本小组考虑到不同区域的生活标准可能存在差距，因此会对运价产生影响，于是将其按照中国标准地理区分（华南，华东，华北，华中，东北，西南，西北）分为七个不同的类别，再将划分好的七个类型处理成哑变量。（包含起始地和目的地）

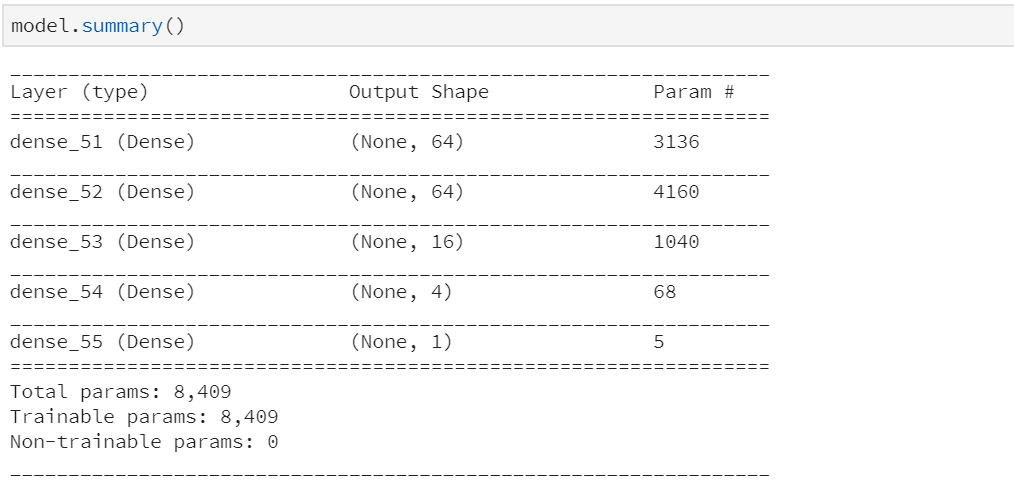
7.对于天气数据，考虑到不同天气和温度可能对运价造成的影响，本小组基于业务分析和数据分析，采用了起始地的平均气温，对其进行了归一化处理，作为特征。

8.对于油价数据，考虑到其直接影响运价，并且不同时期各个地方的油价均存在差别，因此本小组采用了起始地的油价数据，并对其进行了归一化处理看，作为特征。

9.对运价数据进行加0.4176之后取根号操作，使之对称。

### 五、数据建模与结果分析

本小组主要使用了神经网络进行数据建模，进行多次试验和交叉验证后，选择使用三层神经层的全连接神经网络，激活函数为relu。模型具体参数如图五：



图五 神经网络模型参数

根据模型交叉验证结果，模型平均绝对误差稳定在0.16左右，结果较好，由于时间关系，没能更进一步优化模型和调参，如果能进入决赛再进行更进一步工作。