目录

一、 背景介绍.....2
（一） 项目背景.....2
（二） 分析目标.....2
二、 数据说明.....2
（一） 数据基本属性.....2
（二） 数据预处理.....4
三、 数据基本分析.....5
（一） 数据的基本统计量.....5
（二） 菜品销量 Top10 展示.....6
四、 构建菜品推荐模型.....7
（一） 模型算法原理.....7
1. ItemCF 算法原理.....7
2. ItemCF 算法步骤.....7
（二） 数据预处理.....7
（三） 划分训练集和测试集.....8
1. 划分测试集标签与训练集标签.....8
2. 生成“客户 - 菜品”二元矩阵.....8
（四） 模型构建.....8
1. 计算菜品与菜品之间的相似度矩阵.....8
2. 生成推荐列表.....9
（五） 构建模型评价指标.....9
（六） 模型的评估与优化.....9
五、 小结.....11

一、背景介绍

（一）项目背景

民以食为生。消费者对美食永远充满期待，这在一定程度上促进了整个餐饮行业的服务改善和长远发展。

近几年来，餐饮行业增长势头旺盛，发展突飞猛进，餐饮行业餐费收入处于增长趋势，各地餐饮经营数不胜数。同时，餐饮企业也面临着以下问题：（1）同行竞争压力大；（2）对消费者的需求认识不到位，客户忠诚度第；（3）菜品和服务缺乏个性化。为了提高餐饮企业的竞争力，增强餐饮的个性化服务势在必行。

自从大数据融入餐饮行业后，使得传统餐饮行业向智能化、个性化等方向转变的步伐加快。通过用大数据进行数据分析，可以更为便捷地满足更多消费者的需求，给消费者提供更的服务。 大数据应用于餐饮行业，为从事餐饮行业的企业制定发展规划，明确受众群体，改善服务质量，都起到了不可忽视的作用。

餐饮的多元化发展，满足着不同消费者的需求。利用大数据，可以更加深入地了解消费者的需求，为不同的消费者量身定做相应的产品，并提供配套的服务。

（二）分析目标

订单详情表和订单表提供了某餐饮企业一段时间内的餐饮数据，本项目将通过分析这些数据，为顾客推荐个性化菜品，更好地满足消费者的需求。

二、数据说明

（一）数据基本属性

数据来源于订单详情表说明（meal_order_detail.csv）和订单表说明（meal_order_info.csv），meal_order_detai 表的属性名称与含义如表 1 所示。

名称	含义	名称	含义
detail_id	订单详情 ID	place_order_time	用餐时间
order_id	订单 ID	discount_amt	折扣额度
dishes_id	菜品 ID	discount_reason	折扣说明
logicprn_name	类别名称	kick_back	回扣
parent_class_name	父类名称	add_inprice	添加价格
dishes_name	菜品名称	add_info	添加信息
itemis_add	是否为添加菜	bar_code	条形码
counts	数量	picture_file	图片
amounts	销售金额	emp_id	客户 ID
cost	成本		

表 1.　订单详情表说明（meal_order_detail）

meal_order_info 表的属性名称与含义如表 2 所示。

名称	定义	名称	定义
info_id	订单 ID	lock_time	锁单时间
emp_id	客户 ID	cashier_id	收银 ID
number_consumers	消费人数	pc_id	终端 ID
mode	消费方式	order_number	订单号
dining_table_id	桌子 ID	org_id	门店 ID
dining_table_name	桌子名称	print_doc_bill_num	打印 doc 账单的编码
expenditure	消费金额	lock_table_info	桌子关闭信息
dishes_count	总菜品数	order_status	0 未结算；1 结算；2 已锁单
accounts_payable	付费金额	phone	电话
use_start_time	开始时间	name	名字
check_closed	支付结束		

表 2.　订单表说明（meal_order_info）

（二）数据预处理

对于表 meal_order_detail，有些属性对应的数据无实际意义，应该删除掉，即删除 logicprn_name、parent_class_name、cost、discount_amt、discount_reason、kick_back、add_info、bar_code 共八个属性。

有些属性的数据均一样，例如各个订单的 itemis_add 属性的值均为 0，不存在差异性，对数据分析没有帮助，也要去除。即删除 itemis_add、add_inprice 共两个属性。

从属性的实际意义来看，picture_file 对数据分析也没有帮助，因此也去除。

对于表 meal_order_info，同样地，去除数据无意义的属性 mode、check_closed、cashier_id、pc_id、order_number、print_doc_bill_num、lock_table_info 共七个，去除对数据分析无实际的属性 phone、name 共两个。

此外，对于各个订单的属性 order_status，进行了简单的分析，发现 order_status 的信息对应的订单数量如表 3 所示。

数据	订单数量	订单占比
0（未结算）	9	0.0095
1（结算）	933	0.9873
2（已锁单）	3	0.0032

表 3.　order_status 简单分析

显而易见，结算状态的订单的数量远大于未结算和已锁单的订单数量。因此，为了更加便于数据分析的，删除未结算和已锁单的全部订单信息，同时删除 order_status 属性。

最终，两个表格的属性说明如表 4 所示。

订单详情表说明（meal_order_detail）			订单表说明（meal_order_info）		
序号	名称	含义	序号	名称	含义

0	detail_id	订单详情 ID	0	info_id	订单 ID
1	order_id	订单 ID	1	emp_id	客户 ID
2	dishes_id	菜品 ID	2	number_consumers	消费人数
3	dishes_name	菜品名称	3	dining_table_id	桌子 ID
4	counts	数量	4	dining_table_name	桌子名称
5	amounts	销售金额	5	expenditure	消费金额
6	place_order_time	用餐时间	6	dishes_count	总菜品数
7	emp_id	客户 ID	7	accounts_payable	付费金额
			8	use_start_time	开始时间
			9	lock_time	锁单时间
			10	org_id	门店 ID

表 4.　meal_order_detail 与 meal_order_info 的属性说明

三、数据基本分析

（一）数据的基本统计量

数据各属性的基本统计量如表 5 所示。

meal_order_detail 基本统计量						
属性	数量	均值	标准差	最小值	中位数	最大值
detail_id	10037	4712.339344	1747.410959	753	4666	8246
order_id	10037	802.7756302	320.2090319	137	780	1324
dishes_id	10037	609985.155	118.4123984	606000	609983	610072
counts	10037	1.108498555	0.611015942	1	1	10
amounts	10037	44.82136096	35.81543503	1	35	178

emp_id	10037	1207.549766	166.8006911	982	1147	1610
meal_order_info 基本统计量						
属性	数量	均值	标准差	最小值	中位数	最大值
info_id	933	771.5144695	323.4197054	137	747	1324
emp_id	933	1205.079314	167.6237872	982	1147	1610
number_consumers	933	5.201500536	2.360284295	1	5	10
dining_table_id	933	1474.714898	82.93347664	459	1486	1519
dining_table_name	933	1010.725616	8.83004514	1001	1008	1040
expenditure	933	491.4469453	266.6946032	48	451	1314
dishes_count	933	11.82207931	5.633901175	1	11	36
accounts_payable	933	491.4469453	266.6946032	48	451	1314
org_id	933	329.6216506	2.08495599	304	330	330

表 5.　各个属性的基本统计量

（二）菜品销量 Top10 展示

由于收集数据环节中的失误，meal_order_detail 表中的 dishes_name（菜品名称）数据中含有无关字符“\r”和“\n”，需要去除它们。

接下来构建热销度评分指标，计算公式如式[1]所示。

$$\lambda_{\text{热销度评分}} = \frac{Q_i - Q_{min}}{Q_{max} - Q_{min}} \quad [1]$$

其中，$\lambda_{\text{热销度评分}}$ 为某项菜品的热销度评分，值范围在 0 到 1 之间；Q_i 为某项菜品的销售份数；Q_{max} 为最近 30 天内菜品的最大销售份数；Q_{min} 为最近 30 天内菜品的最小销售份数。

考虑到实际情况，把白饭/大碗和白饭/小碗剔除后，销售量最大的为凉拌菠菜，共 269 份；销售量最小的为铁板牛肉，共 3 份。各个菜品的热销度评分如图 1 所示。

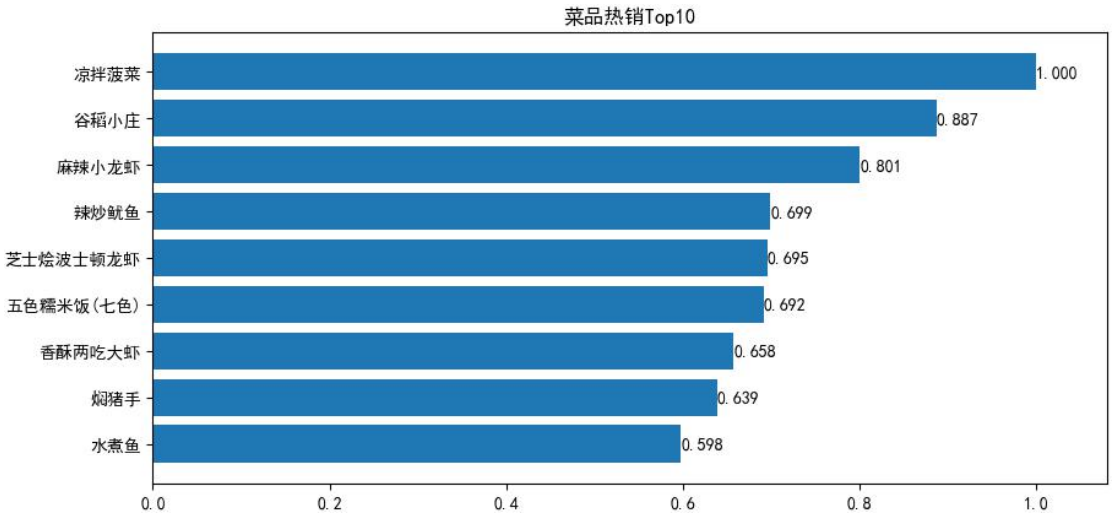

图 1.　　菜品热销 Top10

四、构建菜品推荐模型

（一）模型算法原理

本模型意义在于为客户推荐菜品。由于客户数量大于菜品数量，所以选择基于物品的协同过滤算法（即 ItemCF 算法）。

1．ItemCF 算法原理

基于物品的协同过滤算法就是给用户推荐那些和他们之前喜欢的物品相似的物品，它主要通过分析用户的行为记录计算物品之间的相似度。该算法认为，物品 A 和物品 B 具有很大的相似度是因为喜欢物品 A 的用户大都也喜欢物品 B。

2．ItemCF 算法步骤

① 计算菜品与菜品之间的相似度

② 根据菜品的相似度和客户的历史行为给客户生成推荐表。

（二）数据预处理

为了保持信息量，只保留点了 3 个菜以上的客户，其余客户的订单均删除。共删除 11 个客户信息。

部分 meal_order_detail 表中的订单在 meal_order_info 表中不存在，这些订单均为无效订单，需要删除。

由于本模型的目的是为客户推荐喜欢的菜品，因此菜品名称、和客户 ID 为主要特征信息，即使用 meal_order_detail 表中的 dishes_name 和 emp_id 构建模型，且剔除白饭/大碗和白饭/小碗。

（三）划分训练集和测试集

1．划分测试集标签与训练集标签

将 meal_order_info 表中的 emp_id 去重，并划分 0.2 为训练集标签，0.8 为测试集标签。训练集标签共 372 个，测试集标签共 94 个。

2．生成“客户 - 菜品”二元矩阵

分别使用训练集标签和测试集标签，基于 meal_order_detail 表中的 dishes_name 和 emp_id，生成“客户 - 菜品”二元矩阵 $T_{372\times143}$ 和 $R_{94\times143}$ 。

其中，“0”表示该客户曾经点过这道菜，“1”表示该客户没有点过这道菜。

（四）模型构建

1．计算菜品与菜品之间的相似度矩阵

计算物品间的相似度常用方法有三种，分别是欧几里得距离、皮尔逊系数、Jaccard 相似度。这里使用 Jaccard 相似度，计算公式如式[2]所示。

$$\mathrm{J}(\alpha,\beta)=\frac{\left|U_\alpha\cap U_\beta\right|}{\left|U_\alpha\cup U_\beta\right|}=\frac{\left|U_\alpha\cap U_\beta\right|}{\left|U_\alpha\right|+\left|U_\beta\right|-\left|U_\alpha\cap U_\beta\right|} \qquad [2]$$

其中，$\mathrm{J}(\alpha,\beta)$ 表示菜品 α 与菜品 β 的 Jaccard 相似度，$\left|U_\alpha\cap U_\beta\right|$ 表示共同点过菜品 α 和菜品 β 的客户数量，$\left|U_\alpha\right|$ 表示点过菜品 α 的客户数量，$\left|U_\beta\right|$ 表示点过菜品 β 的客户数量。

利用训练集矩阵 $T_{372\times143}$ 计算得到菜品之间的相似度矩阵，计算公式如式[3]所示。

$$J_{143\times143}=(j_{ij})_{143\times143} \quad [3]$$

其中，i 和 j 分别表示第 i 个菜品和第 j 个菜品，取值范围如式[4]所示。

$$j_{ij}=\begin{cases}J(i,j)\in[0,1) & i\neq j\\ 0 & i=j\end{cases} \quad [4]$$

2．生成推荐列表

首先计算测试集中的客户对所有菜品的感兴趣程度，计算公式如式[5]所示。

$$P_{94\times143}=(p_{ij})_{94\times143}=R_{94\times143}\times J_{143\times143} \quad [5]$$

其中，$R_{94\times143}$ 为测试集“客户 - 菜品”二元矩阵，$J_{143\times143}$ 为菜品相似度矩阵。

接着根据 $P_{94\times143}$ 为测试集客户生成菜品推荐列表。

推荐菜品名单保存到文件 p_dishes_name.csv。

（五）构建模型评价指标

推荐菜品的准确率计算公式如式[6]所示

$$p=\frac{F}{T} \quad [6]$$

其中，F 为正确推荐的菜品数，T 为所有推荐的菜品数。

（六）模型的评估与优化

在生成推荐列表时，不同的生成原则会影响式[6]中的 F 和 T，进而影响模型的准确度。为了尽可能使模型达到最佳准确度，需要计算不同生成原则下的推荐菜品的准确率 p。

设定的生成原则是：推荐各个客户感兴趣程度排名前 n 的菜品。

多次改变 n 的值，能得到不同的 p。n 取 5 到 10 之间的整数。

对于同一个 n，进行三次模型训练，取平均准确率作为 p 值。p 保留 5 位小数。结果如表 6 所示。

n	P（第一次）	P（第二次）	P（第三次）	P（平均值）
5	0.17872	0.21276	0.17872	0.19007
6	0.20390	0.19504	0.17198	0.19031
7	0.21884	0.18541	0.20517	0.20314
8	0.19547	0.22074	0.24202	0.21941
9	0.18439	0.21158	0.20804	0.20134
10	0.18191	0.21914	0.18723	0.19609

表 6.　不同 n 值下的模型准确度

为了更直观地展示 p（平均值）关于 n 的变化趋势，绘画 p（平均值）关于 n 的折线图，如图 2 所示。

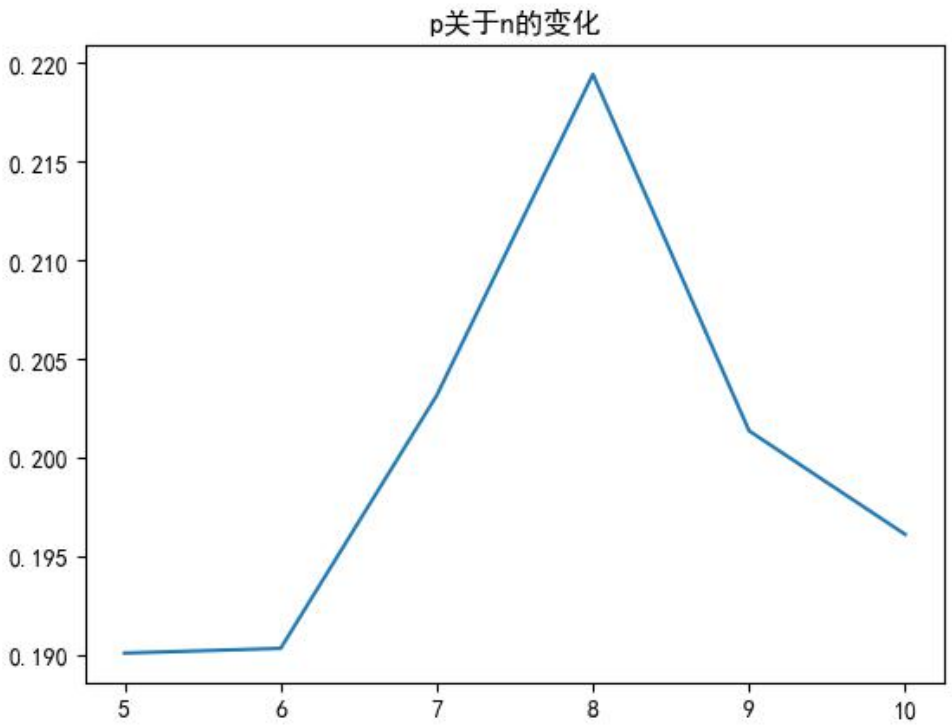

图 2.　p 关于 n 的变化趋势

由图 2 可知，当 n 为 8 时，三次训练的平均准确率 p 值达到最大值 21.941%，故 n 取 8，即向客户推荐感兴趣程度排名前 8 的菜品。

（七）模型的优缺点

本模型的优点是操作简单，训练与测试时间不长，且拥有一定的准确度，适用于简单的菜品推荐，对小餐饮企业来说是不错的选择。

但模型仍存在以下缺点：

① 模型准确度不高

由表 6 可知，准确率最高只能达到 24.202%，平均准确率只有 21.941%。主要是因为用于训练的用户数量小，只有 372 个，若能提供更多时间段的数据，则可以提高菜品推荐的准确度。

② 模型准确度的计算过于粗糙

由于时间和操作的限制，本项目只针对了特定的 n 值进行了 3 次训练，训练次数太少，得到模型准确度存在较大的误差。

五、小结

本项目先简单分析了某餐饮企业一段时间内的餐饮数据，并直观地展示了数据分析的结果，且列出了前 10 个热销菜品，给餐饮企业提供了一定的菜品信息。

再通过某餐饮企业一段时间内的餐饮数据，构建基于物品的协同过滤算法模型，实现了给客户推荐个性化菜品的目标。

从推荐结果来看，模型具有一定的可用性，但仍需进一步的优化。

六、附件说明

a. 代码 data_analysis.py：项目模型代码（可以直接运行，输出模型准确率）
b. 代码 other_code.py：除项目主模型外的其它代码的集合（不能直接运行）
c. 表格 dishes_score.csv：菜品热销度评分计算的结果，用于生成图 1，由 ther_code.py 生成
d. 图片 Figure_1：图 1 的原文件，由 other_code.py 生成
e. 图片 Figure_2：图 2 的原文件，由 other_code.py 生成

f. 表格 info_describe.csv：表 5 的原文件，由 other_code.py 生成

g. 表格 detail_describe.csv：表 5 的原文件，由 other_code.py 生成

h. 表格 p_dishes_name.csv：推荐的菜品列表，由 data_analysis.py 生成