การวิเคราะห์ชุดข้อมูลการใช้จ่ายประจำปีของผู้จัดจำหน่าย

ด้วย Rapidminer

1.Business Understanding

ที่มาของข้อมูล

ชุดข้อมูลนี้เป็นชุดข้อมูล ค่าใช้จ่ายประจำปีของการจัดจำหน่ายขายส่ง ในหน่วยการเงิน (m.u.) หรือ BMD(ดอลลาร์เบอร์มิวดา) ในหมวดหมู่ผลิตภัณฑ์ต่างๆ ในเมือง ของประเทศโปรตุเกต ในปี 2014

แหล่งข้อมูลจาก จากเว็บไซต์ www.archive.ics.uci.edu วันที่ลง ข้อมูล 31-03-2014

ข้อมูลที่ต้องการ

1.ต้องการแบ่งกลุ่มข้อมูล เมื่อแบ่งออกเป็น 2 กลุ่มจะได้ข้อมูลอะไรบ้าง

2.ต้องการทำนายข้อมูล Test ว่าอยู่ ช่องทางการจำหน่ายไหน

3.ต้องการเปรียบเทียบโมเดล ระหว่าง Neural Networks กับ K-NN ว่าโมเดลไหนมีเปอร์เซ็นความแม่นยำสูงกว่ากัน เพื่อหาโมเดลที่เหมาะสมกับข้อมูลชุดนี้

2.Data Understanding

รายละเอียดข้อมูล

มีจำนวณ Instances = 440 และ จำนวณ Attributes = 8

รายการ Attribute

1) FRESH: ผลิตภัณฑ์สด
2) MILK: ผลิตภัณฑ์นม
3) GROCERY: ผลิตภัณฑ์ของชำ
4) FROZEN: ผลิตภัณฑ์แช่แข็ง
5) DETERGENTS_PAPER: ผงซักฟอกและผลิตภัณฑ์กระดาษ
6) DELICATESSEN: ผลิตภัณฑ์อาหารสำเร็จรูป
7)ช่องทางการจำหน่าย(Channel) มี 2 ช่องทาง คือ
   1.Horeca คือ ช่องทางธุรกิจ ประเภทโรงแรม ร้านอาหาร และร้านคาเฟ่
   2.Retail คือ ช่องทางค้าปลีก
8.พื้นที่(Region)
   1. Lisbon คือ เมืองหลวง ลิสบอน ที่ใหญ่อันดับ 1 ของประเทศโปรตุเกต
   2. Oporto คือ เมือง โอพอร์โต ที่ใหญ่อันดับสอง ของประเทศโปรตุเกส
   3. Other Region คือ เมืองอื่นๆ ในประเทศโปรตุเกส

3. Data Preparation

ขั้นตอนการเตรียม Data Preparation

1.เลือกข้อมูลที่ต้องการนำมาวิเคราะห์

2.ตรวจสอบและทำความสะอาดข้อมูล ที่ หาย หรือ มีรูปแบบข้อมูลที่ผิดแปลกไป

A	B	C	D	E	F	G	H
Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Pap	Delicassen
2	3	12669	9656	7561	214	2674	1338
2	3	7057	9810	9568	1762	3293	1776
2	3	6353	8808	7684	2405	3516	7844
1	3	13265	1196	4221	6404	507	1788

Channel	Fresh	Milk	Grocery	Frozen	Detergents_Pap	Delicassen
2	12669	9656	7561	214	2674	1338
2	7057	9810	9568	1762	3293	1776
2	6353	8808	7684	2405	3516	7844

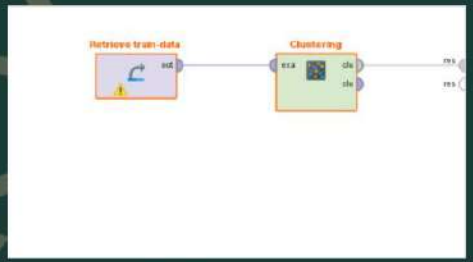
4.Model Building

การใช้โมเดล

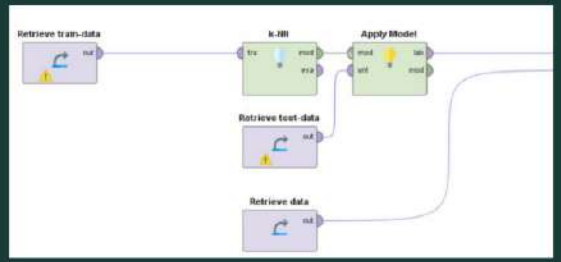
1.K-Means สำหรับการแบ่งกลุ่ม

2.K-NN และ Neural Net สำหรับการทำนายผล

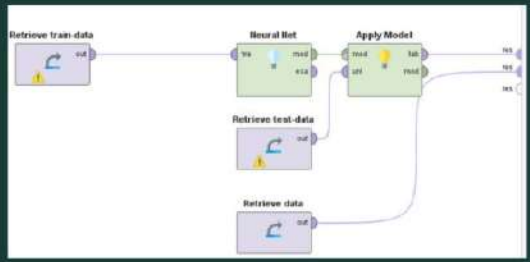
ภาพกระบวนการใช้โมเดล



K-Means



K-NN

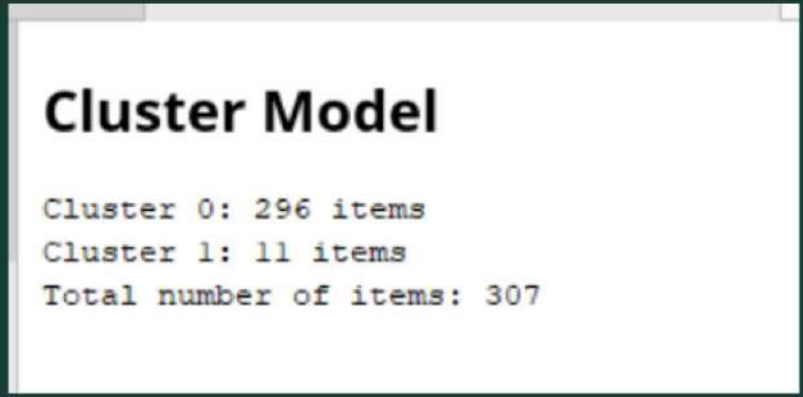


Neural Net

5.Testing and Evaluation

ผลการทดสอบและประเมิน

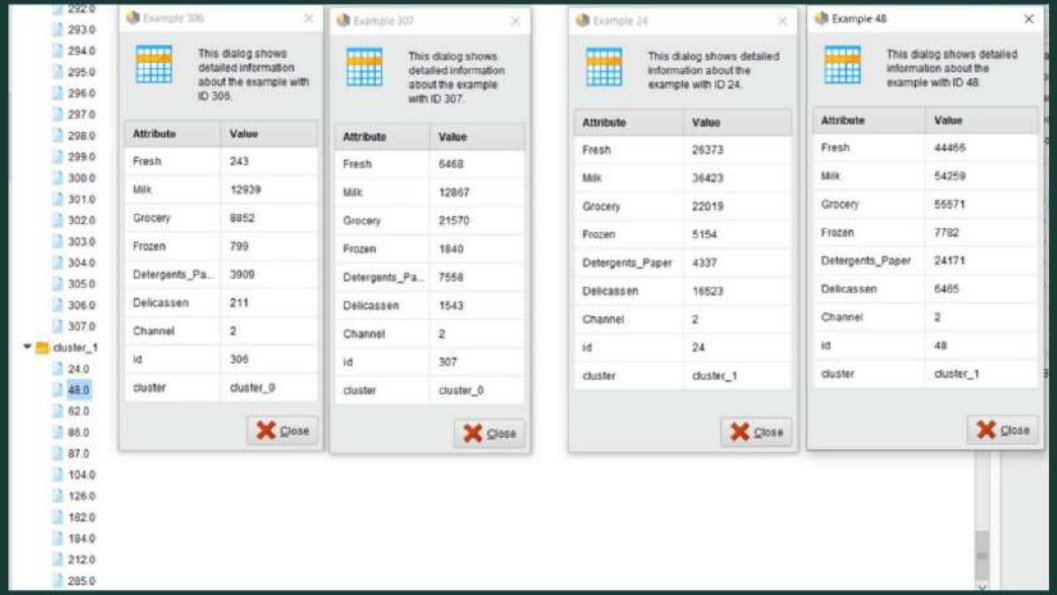
1.ผลการแบ่งกลุ่ม ด้วยโมเดล K-Means โดยกำหนด K=2



จากรูปก็จะพบว่า แบ่งออกได้ 2 กลุ่ม โดย กลุ่ม 0 มีจำนวน 296 Items และกลุ่มที่2 มีจำนวน 307 Items

ผลการทดสอบและประเมิน

1.ผลการแบ่งกลุ่ม ด้วยโมเดล K-Means โดยกำหนด K=2



เมื่อนำข้อมูลจากกลุ่ม 1 example ที่ 306, 307 และ กลุ่ม 2 example ที่ 24, 48 มาวิเคราห์คราวๆก็จะพบว่าข้อมูลในกลุ่มที่ 1 นั้น จะมีค่าเฉลี่ยรวมในแต่ละผลิตภัณฑ์ น้อยกว่า กลุ่ม 2 อย่างเห็นได้ชัด

ผลการทดสอบและประเมิน

2.ผลการทำนายผล ข้อมูล Test ด้วยโมเดล K-NN โดยกำหนด K=3

Open in Turbo Prep Auto Model Filter (133 / 133 examples): all

Row No.	prediction(C...	confidence(2)	confidence(1)	Fresh	Milk	Grocery	Frozen	Detergents_...	Delicassen
1	1	0	1	17327	2374	2842	1149	351	925
2	1	0	1.000	6987	1029	3007	416	257	656
3	2	1	0	918	20655	13567	1465	6846	806
4	1	0	1	7034	1492	2405	12569	299	1117
5	1	0	1	29635	2335	8280	3046	371	117
6	2	1	0	2137	3737	19172	1274	17120	142
7	1	0	1	9784	925	2405	4447	183	297
8	1	0.322	0.678	10617	1795	7647	1483	857	1233
9	2	1	0	1479	14982	11924	662	3891	3508
10	1	0	1	7127	1375	2201	2679	83	1059
11	1	0	1	1182	3088	6114	978	821	1637
12	1	0	1	11800	2713	3558	2121	706	51
13	2	1	0	9759	25071	17645	1128	12408	1625
14	1	0	1	1774	3696	2280	514	275	834
15	1	0	1	9155	1897	5167	2714	228	1113
16	1	0	1	15881	713	3315	3703	1470	229
17	1	0.344	0.656	13360	944	11593	915	1679	573
18	1	0	1	25977	3587	2464	2369	140	1092

รูปที่ 1

Open in Turbo Prep Auto Model Filter (

Row No.	Channel	Fresh	Milk	Grocery	Frozen	Detergents_...	Delicassen
1	1	17327	2374	2842	1149	351	925
2	1	6987	1029	3007	416	257	656
3	2	918	20655	13567	1465	6846	806
4	1	7034	1492	2405	12569	299	1117
5	1	29635	2335	8280	3046	371	117
6	2	2137	3737	19172	1274	17120	142
7	1	9784	925	2405	4447	183	297
8	1	10617	1795	7647	1483	857	1233
9	2	1479	14982	11924	662	3891	3508
10	1	7127	1375	2201	2679	83	1059
11	1	1182	3088	6114	978	821	1637
12	1	11800	2713	3558	2121	706	51
13	2	9759	25071	17645	1128	12408	1625
14	1	1774	3696	2280	514	275	834
15	1	9155	1897	5167	2714	228	1113
16	1	15881	713	3315	3703	1470	229
17	1	13360	944	11593	915	1679	573
18	1	25977	3587	2464	2369	140	1092

รูปที่ 2

จากการใช้ โมเดล K-NN ในการทำนาย หาช่องทางการจำหน่าย ในรูปที่ 1 ก็พบว่าผลการทำนายส่วนใหญ่ตรงกับเฉลย ในรูปที่ 2

ผลการทดสอบและประเมิน

3.ผลกาทำนายผล ข้อมูล Test ด้วยโมเดล Neural Net โดยกำหนด training cycle=900

Open in Turbo Prep Auto Model Filter (133 / 133 examples): all

Row No.	prediction(C...	confidence(2)	confidence(1)	Fresh	Milk	Grocery	Frozen	Detergents_...	Delicassen
22	1	0.002	0.998	16933	2209	3389	7849	210	1534
23	1	0.006	0.994	5113	1486	4583	5127	492	739
24	1	0.010	0.990	9790	1786	5109	3570	182	1043
25	2	0.994	0.006	11223	14881	26839	1234	9606	1102
26	1	0.014	0.986	22321	3216	1447	2208	178	2602
27	2	0.997	0.003	8565	4980	67298	131	38102	1215
28	1	0.001	0.999	16823	928	2743	11559	332	3486
29	2	0.847	0.153	27082	6817	10790	1365	4111	2139
30	1	0.025	0.975	13970	1511	1330	650	146	778
31	1	0.002	0.998	9351	1347	2611	8170	442	868
32	1	0.001	0.999	3	333	7021	15601	15	550
33	1	0.002	0.998	2617	1188	5332	9584	573	1942
34	2	0.967	0.033	381	4025	9670	388	7271	1371
35	2	0.906	0.094	2320	5763	11238	767	5162	2158
36	2	0.805	0.195	255	5758	5923	349	4595	1328
37	2	0.996	0.004	1689	6964	26316	1456	15469	37
38	1	0.011	0.989	3043	1172	1763	2234	217	379
39	2	0.708	0.292	1198	2602	8335	402	3843	303

รูปที่ 1

Open in Turbo Prep Auto Model Filter

Row No.	Channel	Fresh	Milk	Grocery	Frozen	Detergents_...	Delicassen
22	1	16933	2209	3389	7849	210	1534
23	1	5113	1486	4583	5127	492	739
24	1	9790	1786	5109	3570	182	1043
25	2	11223	14881	26839	1234	9606	1102
26	1	22321	3216	1447	2208	178	2602
27	2	8565	4980	67298	131	38102	1215
28	2	16823	928	2743	11559	332	3486
29	2	27082	6817	10790	1365	4111	2139
30	1	13970	1511	1330	650	146	778
31	1	9351	1347	2611	8170	442	868
32	1	3	333	7021	15601	15	550
33	1	2617	1188	5332	9584	573	1942
34	2	381	4025	9670	388	7271	1371
35	2	2320	5763	11238	767	5162	2158
36	1	255	5758	5923	349	4595	1328
37	2	1689	6964	26316	1456	15469	37
38	1	3043	1172	1763	2234	217	379
39	1	1198	2602	8335	402	3843	303

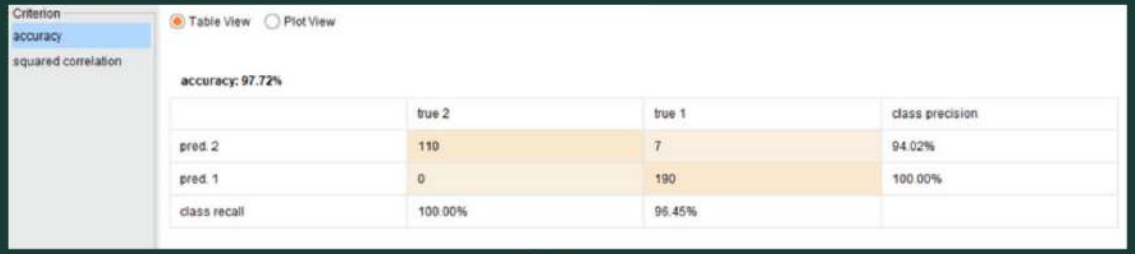
รูปที่ 2

จากการใช้ โมเดล Neural Net ในการทำนาย หาช่องทางการจำหน่าย ในรูปที่ 1 ก็พบว่าผลการทำนายส่วนใหญ่ตรงกับเฉลย ในรูปที่ 2

ผลการทดสอบและประเมิน

4.ผลการหา เปอร์เซ็นต์ความแม่นยำ ของโมเดล K-NN กับ Neural Net

K-NN



รูปที่ 1

รูปที่ 2

จากรูปเมื่อ วัดความแม่นยำแล้วจะได้ค่า Accuracy = 97.72% ในรูปที่ 1 และ squared correlation = 0.907 ในรูปที่ 2 ซึ่งถือว่าเป็นค่าที่สูงมาก

ผลการทดสอบและประเมิน

4.ผลการหา เปอร์เซ็นต์ความแม่นยำ ของโมเดล K-NN กับ Neural Net

Neural Net

รูปที่ 1

รูปที่ 2

จากรูปเมื่อ วัดความแม่นยำแล้วจะได้ค่า Accuracy = 94.14% ในรูปที่ 1 และ squared correlation = 0.761 ในรูปที่ 2 ซึ่งถือว่าเป็นค่าที่กลางๆ แต่ถ้ากำหนด training cycle มากมีค่ามากกว่า 900 ก็จะได้เปอร์เซ็นต์ที่สูงขึ้นแต่แรกกับเวลาในการประมวลผลเยอะขึ้น

ผลการทดสอบและประเมิน

4.ผลการหา เปอร์เซ็นต์ความแม่นยำ ของโมเดล K-NN กับ Neural Net

สรุป

จากการหาเปอร์เซ็นความแม่นยำ ของทั้ง 2 โมเดล พบว่า โมเดล K-NN ให้ค่าเปอร์เซ็นต์ความแม่นยำ สูงกว่า และใช้เวลาในการประมวลผลน้อยกว่า โมเดล Neural Net จึงสรุปได้ว่า โมเดล K-NN เหมาะสมสำหรับใช้กับชุดข้อมูลนี้