โครงงานCRISP-DM

ชื่อโครงงาน สถิติการเช่าจักรยาน

ขั้นตอนที่ 1 [ทำความเข้าใจธุรกิจ]

ไม่ใช่ว่าทุกคนจะมี รถยนต์ เอาไว้ใช้เดินทาง โดยเฉพาะผู้ที่ต้องการเดินทางในระยะทางสั้นๆ ธุรกิจนี้จะตอบโจทย์ผู้ที่ต้องการจะเช่าจักรยานในการเดินทาง แต่เรามีปัญหาที่ว่าหากเราขายราคาเท่ากันตลอดเวลาจะทำให้เราเสียโอกาสที่จะได้กำไรเพิ่มในขณะที่มีความต้องการสูงมิหนำซ้ำหากจักรยานมีไม่มากพอในพื้นที่ก็ทำให้เสียลูกค้าอีก แต่ถ้าหากเราปล่อยเช่าในราคาแพงตลอดเวลาและมีจักรยานกระจุกที่เดียว เมื่อมีความต้องการน้อยเราจะขายไม่ออก และเราก็จะไม่มีจักรยานไปไว้ในที่ๆมีความต้องการมาก ดังนั้นเราต้องการวิธีที่จะทำให้เราสามารถกะจำนวนจักรยานที่จะปล่อยเช่าในช่วงเวลาหนึ่งๆ พร้อมกับสามารถคำนวณราคาที่ควรขายในเวลานั้นๆเพื่อกำไรที่สูงสุด

ขั้นตอนที่ 2 [ทำความเข้าใจข้อมูล]

ข้อมูลแบ่งออกเป็น 2 ไฟล์แบ่งออกเป็นไฟล์ที่แยกเป็นวันและชั่วโมง ณ ที่นี้จะใช้ไฟล์ที่แบ่งชั่วโมงซึ่งจะให้ความละเอียดมากกว่า ข้อมูลประกอบไปด้วย

- instant - บันทึกไอดี/ลำดับของข้อมูล (id)
- dteday - วันที่บันทึกข้อมูล
- season - ฤดูกาล (Spring, Summer, Fall, Winter)
- yr - ปี (0 แทน 2011 และ 1 แทน 2012)
- mnth - เดือน (1 ถึง 12)
- hr - เวลา (0 ถึง 23)
- holiday – วันที่บันทึกเป็นวันหยุดหรือไม่ (อ้างอิงข้อมูลวันหยุดจาก dchh.cd.gov)
- weekday – วันของสัปดาห์ (0 – 6 อาทิตย์ – เสาร์ ตามลำดับ)
- workingday – เป็นวันที่คนไปทำงานหรือไม่ (0 แทน False และ 1 แทน True)
- weathersit – สถาพอากาศ (1 – แจ่มใส 2 – เมฆเยอะ 3 - ฝนตกอ่อน 4 - ฝนตกหนัก)
- temp – อุณหภูมิหลังจากทำการ Normalize แล้วในหน่วยเซลเซียส (หารด้วย 41 max)
- atemp – อุณหภูมิที่รู้สึกหลังจากทำการ Normalize แล้วในหน่วยเซลเซียส (หารด้วย 50 max)
- hum - ความชื้นหลังจากทำการ Normalize แล้วในหน่วยเซลเซียส (หารด้วย 100 max)

> windspeed – ความไวลมหลังทำการ Normalize แล้ว (หารด้วย 67 max)

> casual – จำนวนผู้ใช้ที่ไม่ได้ลงทะเบียนในระบบ

> registered – จำนวนผู้ใช้ที่ลงทะเบียนในระบบแล้ว

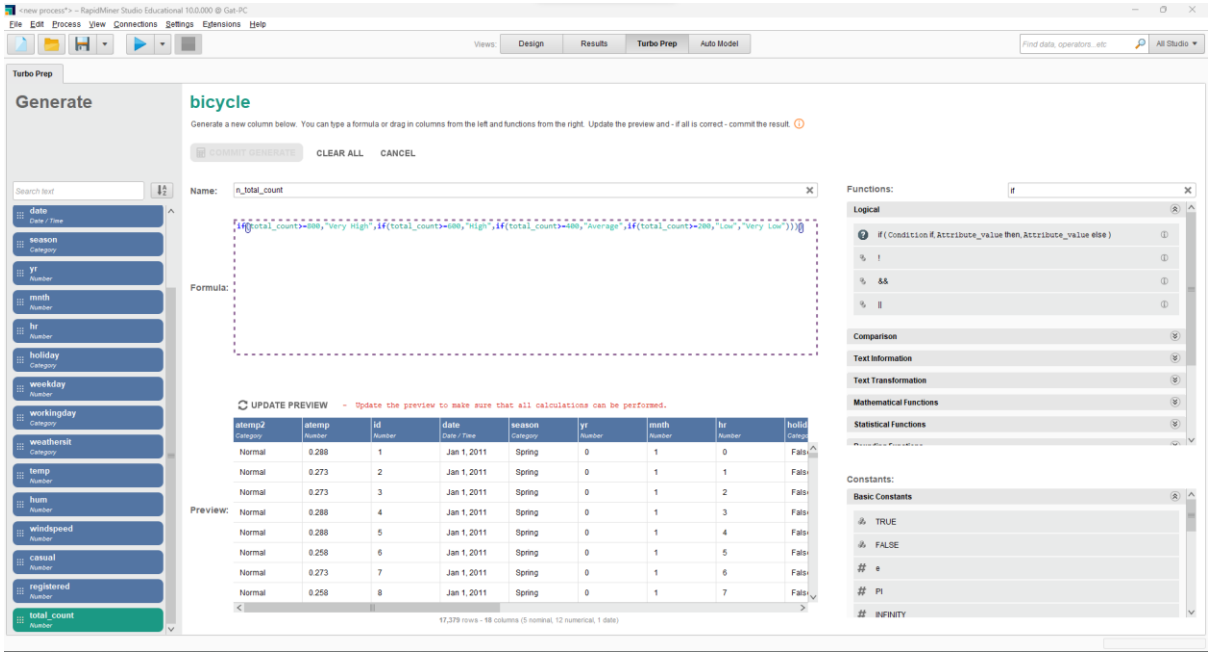
> total_count - จำนวนผู้ใช้รวม (label)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	instant	dteday	season	yr	mnth	hr	holiday	weekday	workingda	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt	
2	1	1/1/2011	1	0	1	0	0	6	0	1	0.24	0.2879	0.81	0	3	13	16	
3	2	1/1/2011	1	0	1	1	0	6	0	1	0.22	0.2727	0.8	0	8	32	40	
4	3	1/1/2011	1	0	1	2	0	6	0	1	0.22	0.2727	0.8	0	5	27	32	
5	4	1/1/2011	1	0	1	3	0	6	0	1	0.24	0.2879	0.75	0	3	10	13	
6	5	1/1/2011	1	0	1	4	0	6	0	1	0.24	0.2879	0.75	0	0	1	1	
7	6	1/1/2011	1	0	1	5	0	6	0	2	0.24	0.2576	0.75	0.0896	0	1	1	
8	7	1/1/2011	1	0	1	6	0	6	0	1	0.22	0.2727	0.8	0	2	0	2	
9	8	1/1/2011	1	0	1	7	0	6	0	1	0.2	0.2576	0.86	0	1	2	3	
10	9	1/1/2011	1	0	1	8	0	6	0	1	0.24	0.2879	0.75	0	1	7	8	
11	10	1/1/2011	1	0	1	9	0	6	0	1	0.32	0.3485	0.76	0	8	6	14	
12	11	1/1/2011	1	0	1	10	0	6	0	1	0.38	0.3939	0.76	0.2537	12	24	36	
13	12	1/1/2011	1	0	1	11	0	6	0	1	0.36	0.3333	0.81	0.2836	26	30	56	
14	13	1/1/2011	1	0	1	12	0	6	0	1	0.42	0.4242	0.77	0.2836	29	55	84	
15	14	1/1/2011	1	0	1	13	0	6	0	2	0.46	0.4545	0.72	0.2985	47	47	94	
16	15	1/1/2011	1	0	1	14	0	6	0	2	0.46	0.4545	0.72	0.2836	35	71	106	
17	16	1/1/2011	1	0	1	15	0	6	0	2	0.44	0.4394	0.77	0.2985	40	70	110	
18	17	1/1/2011	1	0	1	16	0	6	0	2	0.42	0.4242	0.82	0.2985	41	52	93	
19	18	1/1/2011	1	0	1	17	0	6	0	2	0.44	0.4394	0.82	0.2836	15	52	67	
20	19	1/1/2011	1	0	1	18	0	6	0	3	0.42	0.4242	0.88	0.2537	9	26	35	
21	20	1/1/2011	1	0	1	19	0	6	0	3	0.42	0.4242	0.88	0.2537	6	31	37	
22	21	1/1/2011	1	0	1	20	0	6	0	2	0.4	0.4091	0.87	0.2537	11	25	36	
23	22	1/1/2011	1	0	1	21	0	6	0	2	0.4	0.4091	0.87	0.194	3	31	34	
24	23	1/1/2011	1	0	1	22	0	6	0	2	0.4	0.4091	0.94	0.2239	11	17	28	
25	24	1/1/2011	1	0	1	23	0	6	0	2	0.46	0.4545	0.88	0.2985	15	24	39	
26	25	2/1/2011	1	0	1	0	0	0	0	2	0.46	0.4545	0.88	0.2985	4	13	17	
27	26	2/1/2011	1	0	1	1	0	0	0	2	0.44	0.4394	0.94	0.2537	1	16	17	
28	27	2/1/2011	1	0	1	2	0	0	0	2	0.42	0.4242	1	0.2836	1	8	9	
29	28	2/1/2011	1	0	1	3	0	0	0	2	0.46	0.4545	0.94	0.194	2	4	6	

ตัวอย่างข้อมูลดิบก่อนนำมาทำความสะอาด

ขั้นตอนที่ 3 [ทำความสะอาดข้อมูล]

เราต้องเริ่มจากการจัดกลุ่มเป้าหมายของเรา ซึ่งนั้นก็คือจำนวนผู้ใช้รวมทั้งหมด เราจะแปลงจากเลขเพียวๆ ออกเป็นกลุ่มๆ ได้ 5 กลุ่ม นั้นก็คือ ผู้ใช้สูงมาก(Very High), มาก(High), ปานกลาง(Average), น้อย(Low) และ น้อยมาก(Very Low) แทนที่จะใช้เลขไปเลยจะช่วยให้โมเดลเราคาดคะเนได้แม่นกว่าและนำไปใช้จริงได้ง่ายกว่า

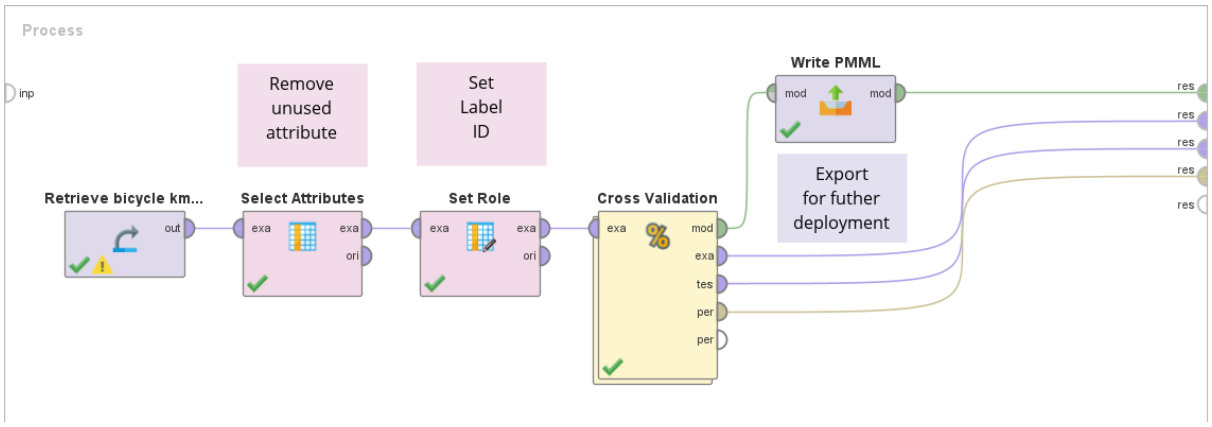


การจัดหมู่ข้อมูล ให้เป็น 5 กลุ่ม เพื่อให้โมเดลมีประสิทธิภาพและใช้จริงได้ง่ายขึ้น

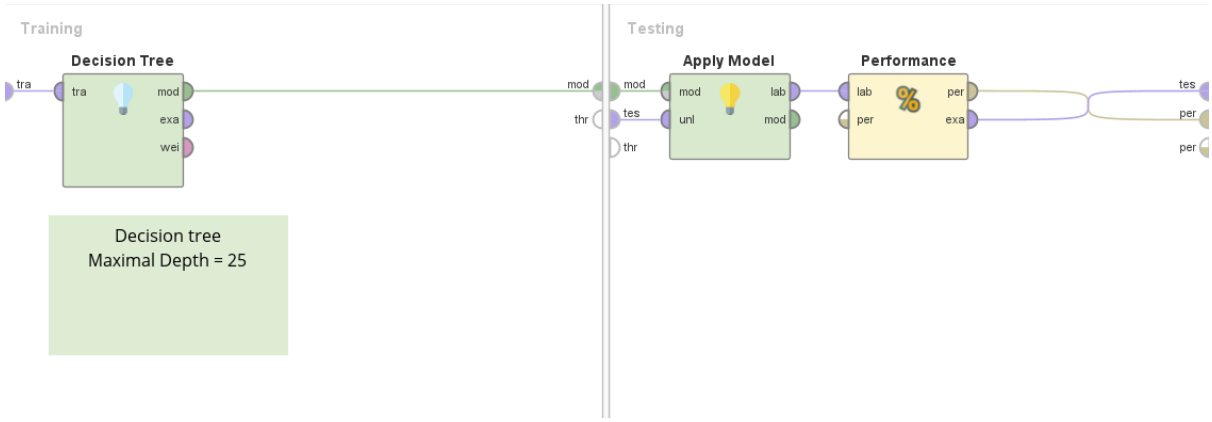
หลังจากนั้นเราจะตัดส่วนที่ไม่ได้ใช้ ได้แก่ วันที่ ปี เนื่องจากมีอัตราการผันผวนสูง และความคงที่ต่ำ ตัดจำนวนผู้ใช้ที่สมัครและยังไม่ได้สมัครออก เนื่องจากเราต้องการจะนับผู้ใช้ทั้งหมดรวมกันเรียบร้อยแล้ว

ขั้นตอนที่ 4 [ทำโมเดล]

เราเลือกใช้ Decision Tree ในการทำโมเดลเนื่องจากพบว่าโมเดล Decision Tree มีอัตราความแม่นยำค่อนข้างมากและใช้เวลาในการประมวลผลน้อยที่สุด



Process การทำโมเดล Decision Tree ภายใน RapidMiner

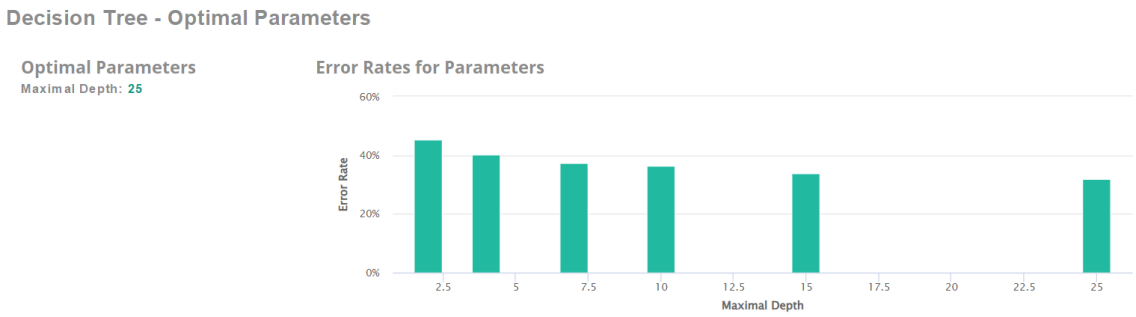


หลังจากได้โมเดลแล้วเราจะส่งออกโมเดลไปเป็นไฟล์ PMML ซึ่งเป็นไฟล์ที่เก็บโครงสร้างของโมเดลเอาไว้เพื่อเตรียมนำไปใช้ในการ Deploy ด้วย python ในขั้นที่ 6

ขั้นตอนที่ 5 [ทำความเข้าใจผลลัพธ์และปรับปรุง]

ในตอนแรกที่ได้ทดลองนั้น พบว่าโมเดล Random Forest มีความแม่นยำมากกว่า Decision Tree เล็กน้อย แต่หลังจากการปรับจูน Maximal Depth ของทั้งสองโมเดลเมื่อ Random Forest มีจำนวนต้นไม้มากกว่า 20 และมี Maximal Depth มากกว่า 7 ซึ่งทำให้ความแม่นยำจะค่อยๆ ตกลงไป ในขณะที่ Decision Tree เมื่อปรับจูน Maximal Depth ไปที่ 25 แล้วมีความแม่นยำสูงกว่า Random Forest โดยวัดจาก Parameter ที่ดีที่สุดของทั้งคู่ แถมยังใช้เวลาในการประมวลผลต่างกันมากจึงเลือกใช้ Decision Tree

Parameter ที่ดีที่สุดของ Random Forest มี error rate อยู่ที่ประมาณ 39%



Parameter ที่ดีที่สุดของ Decision Tree มี error rate อยู่ที่ประมาณ 31%

ขั้นตอนที่ 6 [การนำไปใช้จริง]

หลังจากที่เราได้ model ที่มี accuracy อยู่ประมาณ 70% เราคิดว่า model พร้อมที่จะนำไปพัฒนาต่อเป็น web app V1 ให้ได้ลองใช้งานแล้ว เราจึงนำ upload model ของเราขึ้นไปบนระบบ streamlit เพื่อให้ผู้ใช้ได้ทดลองใช้จริง โดยผู้ใช้สามารถเลือกได้ 2 แบบ คือใช้ข้อมูล real time ของสถานที่นั้นๆ ที่ดึงมาจาก API ของ OpenWeather มาเข้า model เพื่อทำนาย หรืออีกแบบนึงคือ ให้ผู้ใช้กรอกข้อมูลแบบ manual เพื่อนำข้อมูลนั้นมาทำนายได้เช่นกัน โดยผู้ใช้สามารถเอาผลที่ได้จากการทำนายมาลองคำนวณรายได้ วางแผนการเดินทาง หรือแม้แต่การทดลองวางแผนเศรษฐกิจ ในขณะเดียวกัน ผู้ประกอบการสามารถใช้ข้อมูลพยากรณ์สภาพอากาศล่วงหน้าเพื่อคำนวณว่าในแต่ละวันจะเอาจักรยานส่วนมากไปไว้ไหน และเลือกพื้นที่ที่คาดคะเนว่ามีผู้ใช้งานน้อยเพื่อจะจัดโปรโมชั่นพิเศษเพื่อเรียกลูกค้า ไม่ว่าจะเป็น ลดราคา; เหมาจักรยานหลายคันในราคาต่ำ เป็นต้น

ตัวอย่างหน้าเว็บ streamlit :D