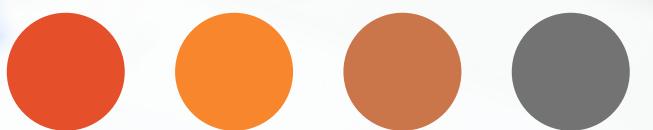


CUSTOMER LIFETIME VALUE

Puphat Maneechan



Business Understanding

สำหรับธุรกิจประเภท Retail ค่อนข้างมีความผันผวนในการซื้อขายในระดับรายลูกค้า พฤติกรรมการซื้อขายของลูกค้ามีความไม่แน่นอนสูง ดังนั้นในการทำโน๊เดลมาวิเคราะห์เพื่อ ช่วยให้การทำธุรกิจมีประสิทธิภาพมากขึ้นจึงควรกำหนด Period ใน การวิเคราะห์ที่ไม่ยาวนาน เกินไป

ในมุมมองของผู้บริหารย่อมอยากรู้ว่า ในอนาคตลูกค้าแต่ละคนจะจ่ายเงินให้เราเป็น ปริมาณเท่าไร เพื่อประมาณการรายได้ที่เราจะได้ในอนาคต ซึ่งจะทำให้การตัดสินใจและการ วางแผนในการดำเนินการต่างๆ เช่น การทำ Marketing Campaign Promotion ต่างๆ มีประสิทธิภาพมากขึ้น สามารถควบคุมต้นทุนได้ดีขึ้น กำหนดกลยุทธ์ในการทำธุรกิจได้ดีขึ้น จึงเป็นที่มาของการ ทำโปรเจค Customer Lifetime Value

Project Objective

สำหรับโปรเจคนี้ คือการคำนวณ Customer Lifetime value ของลูกค้าบน E-commerce Retail โดยมีการวิเคราะห์ คือ

- คำนวณว่าในอีก 3 เดือนข้างหน้า ลูกค้าแต่ละคนจะจ่ายเงิน ซื้อสินค้ากับเราเท่ากันกี่บาท โดยใช้ข้อมูลย้อนหลัง 3 เดือน เพื่อประเมิน Revenue ของเราในอนาคตอีก 3 เดือนข้างหน้าและนำมาช่วยในการวางแผนการทำการตลาดต่างๆ



Table of Contents

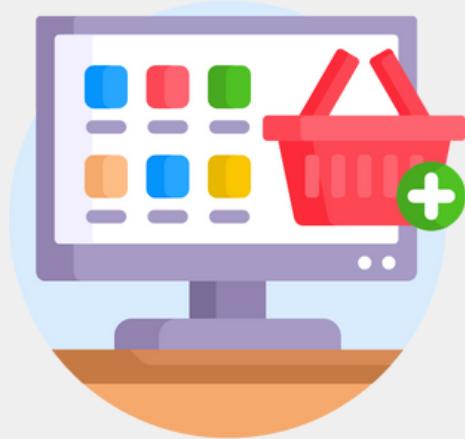
- Project Overview
- Input and Prepare Data
- Regression Model
- Business Impact & Strategies



ส่วนที่ 1 Project Overview



Project Overview



Data Set: Online Retail รูปแบบ Transaction sales in 2011 from kaggle

Unit of analysis: Customer level

Model

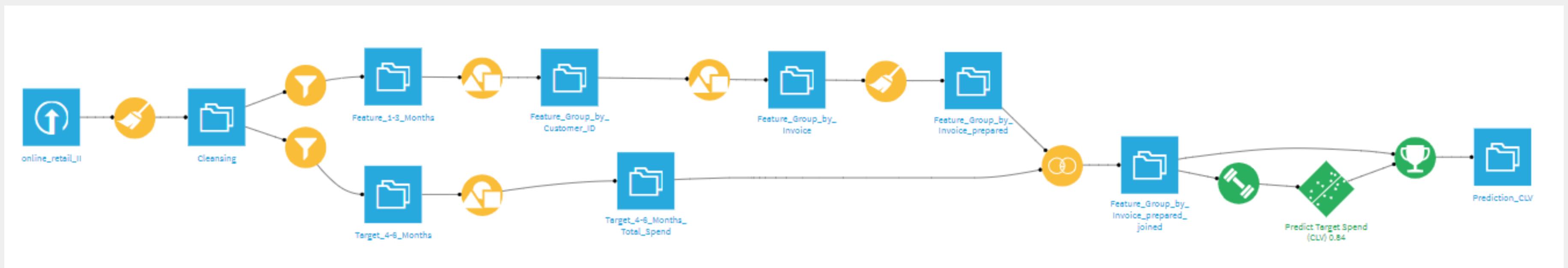
- Regression - คำนวณ Customer Lifetime Value ในอีก 3 เดือนข้างหน้า

Timeline ที่เราใช้จะแบ่งเป็น เดือนที่ 1-3 สำหรับเตรียมตัวแพร X เพื่อไป
คำนวณ เดือนที่ 4-6 เป็นตัวแพร Y



Project Overview

Flow กระบวนการกึ่ง自动化的 Customer Lifetime Value Prediction



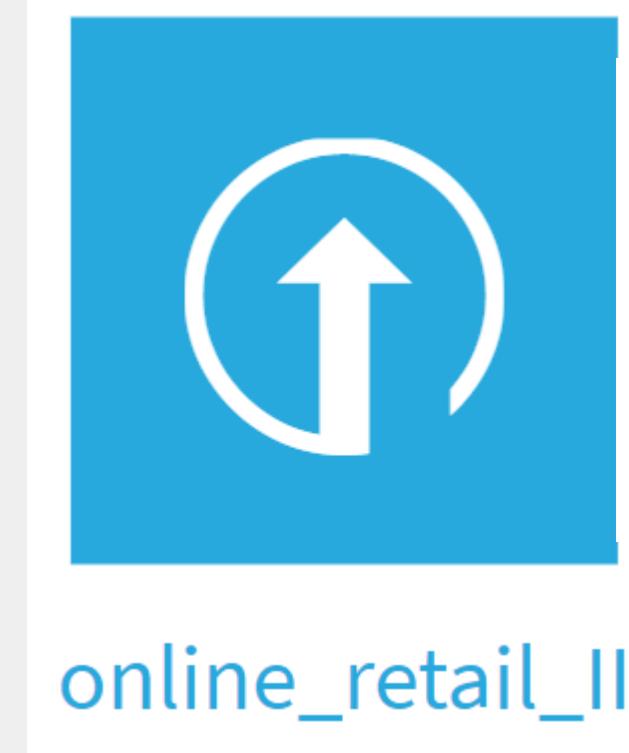


ส่วนที่ 2

Input and Prepare Data



Input Data Set



online_retail_II

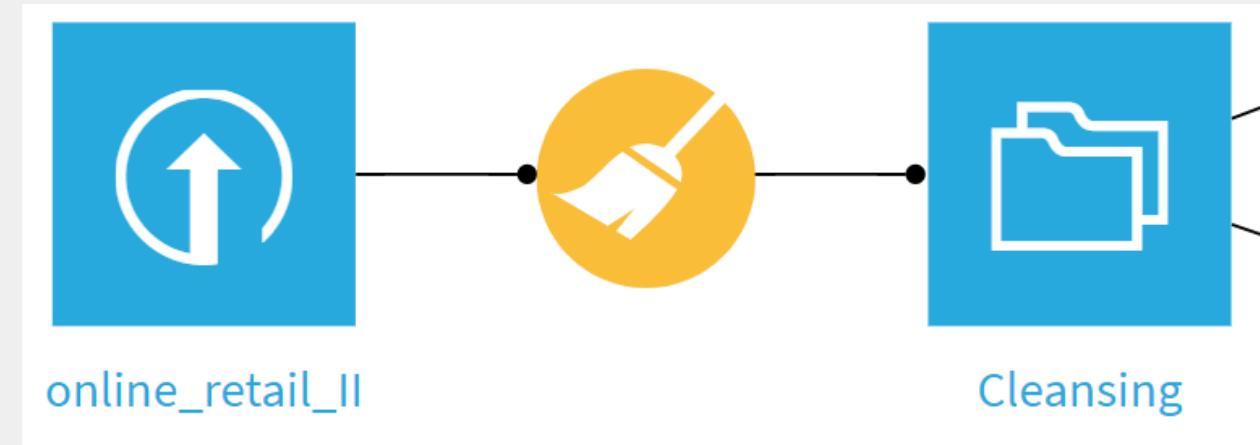
ข้อมูลที่นำเข้าเป็น
Transaction การซื้อสินค้า
ต่างๆของลูกค้า

COLUMN COUNT	FILE COUNT	SIZE	RECORD COUNT
8	1	22.03 MB	541 910
2022-07-12 16:18	2022-07-12 16:18	2022-07-12 16:18	2022-07-12 16:18

ตัวอย่างข้อมูลภายใน

Invoice	StockCode	Description	Quantity	InvoiceDate	Price	Customer ID	Country
string	string	string	string	string	string	string	string
Integer	Integer	Natural lang.	Integer	Text	Decimal	Integer	Country
537688	22909	SET OF 20 VINTAGE CHRISTMAS NAPKINS	12	12/8/10 10:35	0.85	17406	United Kingdom
537688	84406B	CREAM CUPID HEARTS COAT HANGER	32	12/8/10 10:35	2.75	17406	United Kingdom
537688	21166	COOK WITH WINE METAL SIGN	12	12/8/10 10:35	1.95	17406	United Kingdom
537688	21175	GIN + TONIC DIET METAL SIGN	24	12/8/10 10:35	2.1	17406	United Kingdom
537688	85152	HAND OVER THE CHOCOLATE SIGN	12	12/8/10 10:35	2.1	17406	United Kingdom
537688	21174	POTTERING IN THE SHED METAL SIGN	12	12/8/10 10:35	1.95	17406	United Kingdom
537688	21181	PLEASE ONE PERSON METAL SIGN	12	12/8/10 10:35	2.1	17406	United Kingdom
537688	21903	MAN FLU METAL SIGN	12	12/8/10 10:35	2.1	17406	United Kingdom
537688	21908	CHOCOLATE THIS WAY METAL SIGN	12	12/8/10 10:35	2.1	17406	United Kingdom
537688	85150	LADIES & GENTLEMEN METAL SIGN	6	12/8/10 10:35	2.55	17406	United Kingdom

Prepare Data



Remove rows where **Invoice** is not a valid **Integer**
- 100

Remove rows with **Empty values in Customer ID**
- 2291

Keep rows where **Country** is **United Kingdom**
- 569

Move columns **Customer ID, InvoiceDate** at beginning

Parse date in **InvoiceDate**
7040

Create column **Total Spend** with formula
Quantity * Price
7040

Extract date components from
InvoiceDate_parsed
7040

Create column **Return Item** with formula
if(Quantity < 0, 1, 0)
7040

ตัวอย่างข้อมูลภายใน

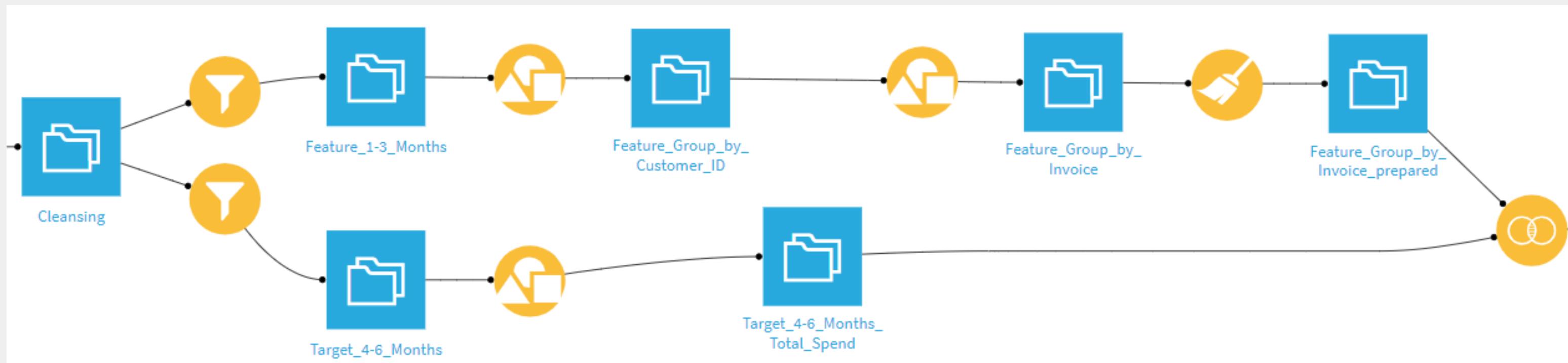
Customer ID	InvoiceDate	InvoiceDate_parsed	Month	Invoice	StockCode	Description	Quantity	Return Item	Price	Total Spend	Country
bigint	string	date	bigint	string	string	string	bigint	bigint	double	double	string
Integer	Date (unparsed)	Date	Integer	Integer	Text	Natural lang.	Integer	Integer	Decimal	Decimal	Country
17850	12/1/10 8:26	2010-12-01T08:26:00.000Z	12	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	0	2.55	15.299999999...	United Kingdom
17850	12/1/10 8:26	2010-12-01T08:26:00.000Z	12	536365	71053	WHITE METAL LANTERN	6	0	3.39	20.34	United Kingdom
17850	12/1/10 8:26	2010-12-01T08:26:00.000Z	12	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	0	2.75	22.0	United Kingdom
17850	12/1/10 8:26	2010-12-01T08:26:00.000Z	12	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	0	3.39	20.34	United Kingdom
17850	12/1/10 8:26	2010-12-01T08:26:00.000Z	12	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	0	3.39	20.34	United Kingdom
17850	12/1/10 8:26	2010-12-01T08:26:00.000Z	12	536365	22752	SET 7 BABUSHKA NESTING BOXES	2	0	7.65	15.3	United Kingdom
17850	12/1/10 8:26	2010-12-01T08:26:00.000Z	12	536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	0	4.25	25.5	United Kingdom
17850	12/1/10 8:28	2010-12-01T08:28:00.000Z	12	536366	22633	HAND WARMER UNION JACK	6	0	1.85	11.100000000...	United Kingdom
17850	12/1/10 8:28	2010-12-01T08:28:00.000Z	12	536366	22632	HAND WARMER RED POLKA DOT	6	0	1.85	11.100000000...	United Kingdom
13047	12/1/10 8:34	2010-12-01T08:34:00.000Z	12	536368	22960	JAM MAKING SET WITH JARS	6	0	4.25	25.5	United Kingdom
13047	12/1/10 8:34	2010-12-01T08:34:00.000Z	12	536368	22913	RED COAT RACK PARIS FASHION	3	0	4.95	14.850000000...	United Kingdom
13047	12/1/10 8:34	2010-12-01T08:34:00.000Z	12	536368	22912	YELLOW COAT RACK PARIS FASHION	3	0	4.95	14.850000000...	United Kingdom
13047	12/1/10 8:34	2010-12-01T08:34:00.000Z	12	536368	22914	BLUE COAT RACK PARIS FASHION	3	0	4.95	14.850000000...	United Kingdom

Prepare Data ดังนี้

- Parse Date
- Remove empty row
- Keep only United Kingdom only เพราะว่ามีปริมาณถึง 90 % ของ Data ทั้งหมด
- สร้าง Total Spend จาก Quantity * Price
- สร้าง Return Item จาก Quantity < 0

Prepare Data

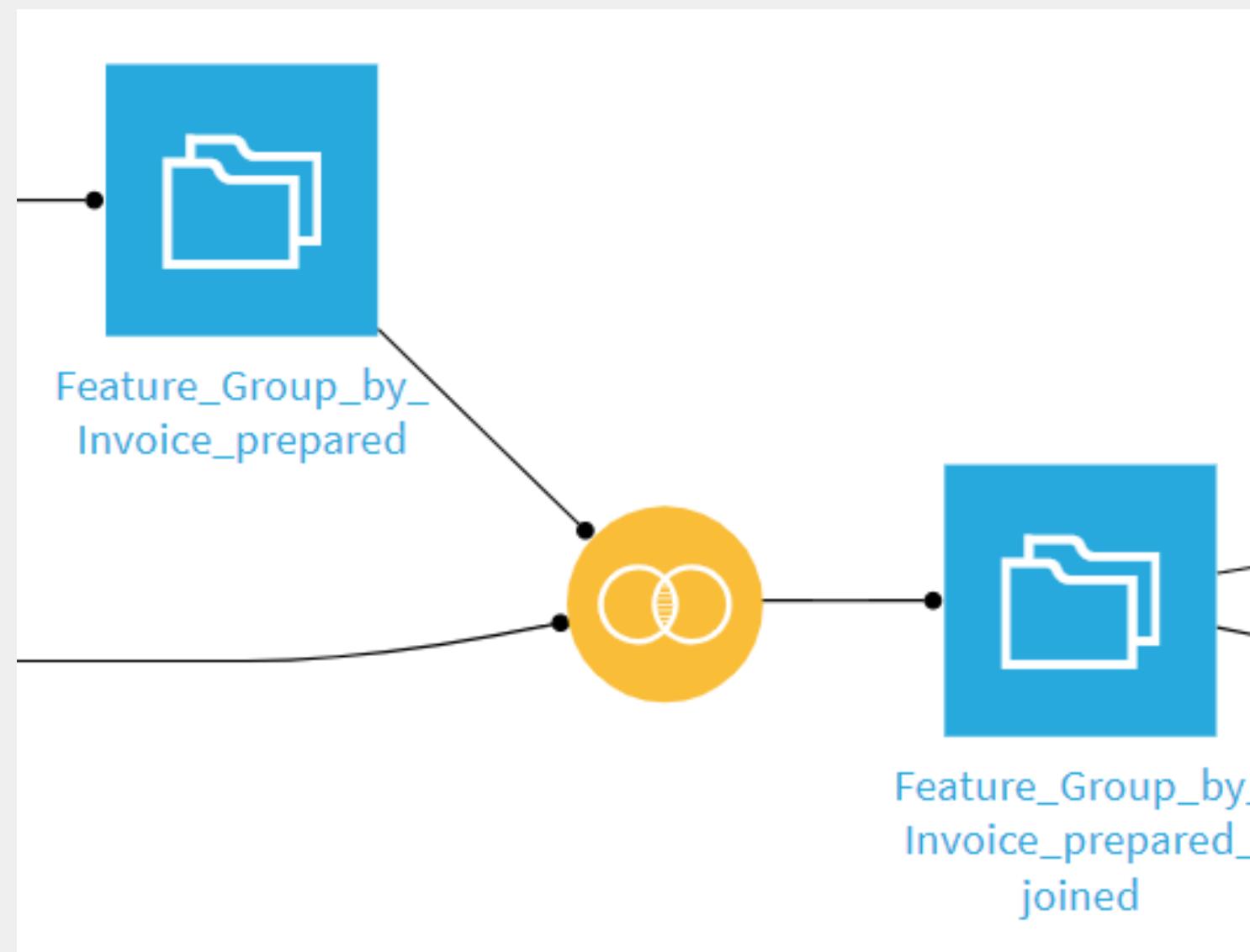
สายบนจะเป็นการสร้าง Feature ตัวแปร X โดยมีการ Filter เลือกเฉพาะเดือนที่ 1-3 และนำมาทำการ Group by จนได้เป็น Customer Single View และนำไปทำการ Prepare ขั้นตอนสุดท้ายเพื่อเตรียม Feature เพิ่ม



สายล่างจะเป็นการสร้าง Target ตัวแปร Y โดยเริ่มที่การ Filter เลือกเฉพาะเดือนที่ 4-6 และต่อมาทำการ Group by Customer ให้ได้เป็น Customer Single View ที่มียอด Total Spend ของเดือนที่ 4-6

Prepare Data

นำ Feature ตัวแปร X กับ Target มา Join กันด้วยวิธี Left Join โดยให้ผู้ Feature เป็นผู้ Left เพื่อเตรียม Data Set ไป Train Model โดยใช้ Customer ID เป็นตัวเชื่อม



Prepare Data

ຕັ້ງອ່າງ Data Set ອີ່ Prepare ເສື້ອເຮືອບຮ້ອຍພຽມ Train Model ມີກົໍ່າມັດ 22 Column

Customer ID	InvoiceDate_parsed_first_first	holiday_Firstbank	holiday_Firstweekend	InvoiceDate_parsed_first_last	holiday_Lastbank	holiday_Lastweekend	Customer Life	Frequency	Mean Time Between Purchase	Recency	DateDiff
bigint Integer	date Date	boolean Boolean	boolean Boolean	date Date	boolean Boolean	boolean Boolean	bigint Integer	bigint Decimal	double	bigint Integer	bigint Integer
12346	2011-01-18T10:01:00.000Z	false	false	2011-01-18T10:01:00.000Z	false	false	71	1	71.0	71	0
12747	2011-01-20T14:01:00.000Z	false	false	2011-03-01T14:53:00.000Z	false	false	69	2	34.5	29	40
12748	2011-01-05T16:06:00.000Z	false	false	2011-03-25T14:45:00.000Z	false	false	84	21	4.0	5	78
12820	2011-01-17T12:34:00.000Z	true	false	2011-01-17T12:34:00.000Z	true	false	72	1	72.0	72	0
12823	2011-02-16T12:15:00.000Z	false	false	2011-03-30T10:36:00.000Z	false	false	42	3	14.0	0	41
12826	2011-01-19T12:52:00.000Z	false	false	2011-01-27T12:06:00.000Z	false	false	70	2	35.0	62	7
12829	2011-01-07T11:13:00.000Z	false	false	2011-01-07T11:13:00.000Z	false	false	82	1	82.0	82	0
12831	2011-03-22T13:02:00.000Z	false	false	2011-03-22T13:02:00.000Z	false	false	8	1	8.0	8	0
12834	2011-03-02T09:49:00.000Z	false	false	2011-03-02T09:49:00.000Z	false	false	28	1	28.0	28	0
12836	2011-02-01T11:19:00.000Z	false	false	2011-02-01T11:19:00.000Z	false	false	57	1	57.0	57	0



Mean Time Between Invoice	Avg Quantity	Total Quantity	Avg Spend	Total Spend	Total Spend stddev	Total Spend first	Total Spend last	Changing Spend	Target Spend
double Decimal	double Decimal	bigint Integer	double Decimal	double Decimal	double Decimal	double Decimal	double Decimal	double Decimal	double Decimal
0.0	74215.0	74215	77183.6	77183.6	0.0	77183.6	77183.6	1.0	1.0
20.0	117.0	234	43.25316666666667	613.82	5.473006486383884	303.04	310.78000000000003	1.0	1147.61
3.7142857142857144	63.7619047619...	1339	9.062533911528353	1987.7799999999995	80.88755778710285	7.8	275.24999999999983	35.0	5341.12999999...
0.0	146.0	146	15.49636363636364	170.46000000000004	0.0	170.46000000000004	170.46000000000004	1.0	
13.666666666666666	43.3333333333...	130	331.5	994.5	116.85568022137392	229.5	459.0	2.0	
3.5	266.5	533	13.50958333333332	542.0999999999999	82.49107709322261	212.71999999999997	329.37999999999994	2.0	121.52
0.0	251.0	251	34.54166666666664	207.25	0.0	207.25	207.25	1.0	
0.0	135.0	135	23.89444444444446	215.05	0.0	215.05	215.05	1.0	



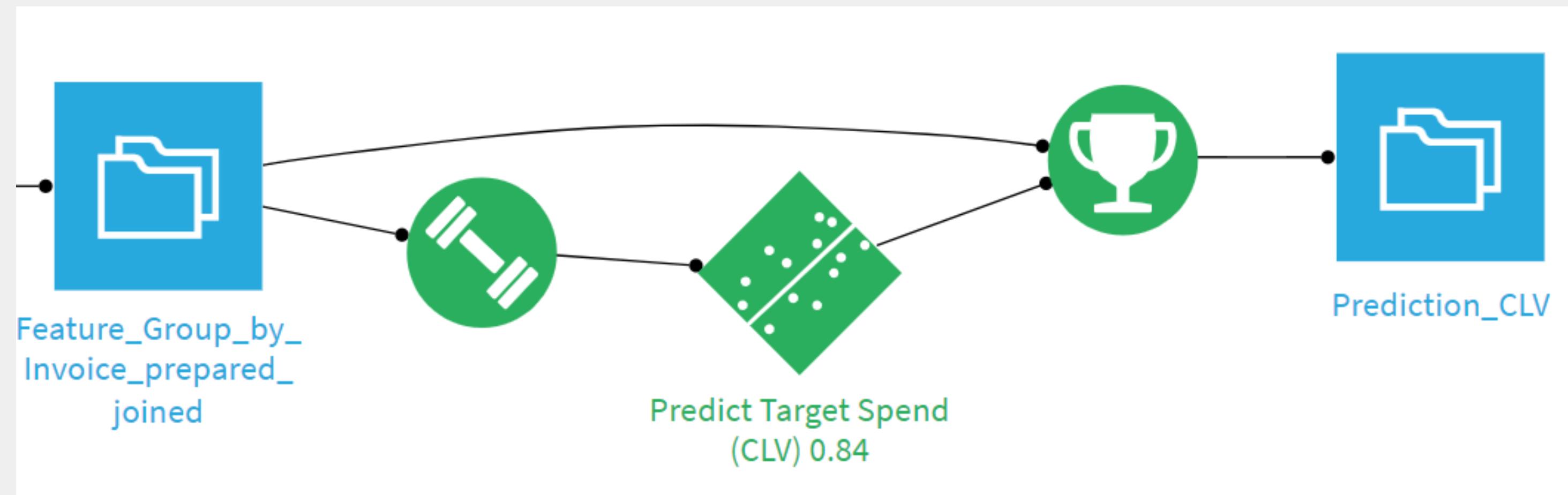
**Changing Spend = Total Spend last / Total Spend first



ส่วนที่ 3 Regression Model



Training Model



Preprocessing

Fill Empty Row ด้วยค่า 0

Fill empty cells of Target Spend with '0'

658

Column single | multiple | pattern | all

Target Spend

Value to fill with

0

x²

+ ADD A NEW STEP

Total Spend last	Changing Spend	Target Spend
Decimal	Decimal	Decimal
77183.6	1.0	0
310.78000000000003	1.0	1147.61
275.24999999999983	35.0	5341.12999999...
170.46000000000004	1.0	0
459.0	2.0	0
329.37999999999994	2.0	121.52
207.25	1.0	0
215.05	1.0	0
312.38	1.0	0

Total Spend last	Changing Spend	Target Spend
Decimal	Decimal	Decimal
77183.6	1.0	0
310.78000000000003	1.0	1147.61
275.24999999999983	35.0	5341.12999999...
170.46000000000004	1.0	0
459.0	2.0	0
329.37999999999994	2.0	121.52
207.25	1.0	0
215.05	1.0	0
312.38	1.0	0

Design Algorithm

Target

Prediction type: Regression

Target: Target Spend

RE-DETECT SETTINGS

Partitioned models

Partitioning: Not available: input dataset is not partitioned.

Target ที่เราจะทำนายคือ Column Target Spend ที่เป็นยอดซื้อของลูกค้าแต่ละคนในช่วง เดือนที่ 4-6

Train / test set for final evaluation

Policy: Split the dataset

Time ordering: Enabled (OFF)

Sampling & Splitting

If your dataset does not fit in your RAM, you may want to subsample the set on which splitting will be performed.

Sampling method: First records

Nb. records: 100000

Split: Randomly For more advanced splitting, use a split recipe, and then use "Explicit extracts from two datasets" policy

K-fold cross-test: Gives error margins on metrics, but strongly increases training time

Train ratio: 0.8 Approximate proportion of the sample that goes to the train set. The rest goes to the test set

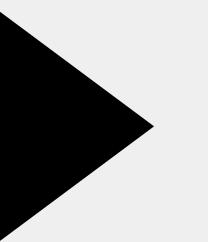
Random seed: 1337 Using a fixed random seed allows for reproducible result

Features Handling

<input type="checkbox"/> # Customer ID Rejected, unique ID	<input type="radio"/> OFF
<input type="checkbox"/> # InvoiceDate_parsed_first_first Reject	<input type="radio"/> OFF
<input type="checkbox"/> A holiday_Firstbank Dummy encoding	<input type="radio"/> ON
<input type="checkbox"/> A holiday_Firstweekend Dummy encoding	<input type="radio"/> ON
<input type="checkbox"/> # InvoiceDate_parsed_first_last Reject	<input type="radio"/> OFF
<input type="checkbox"/> A holiday_Lastbank Dummy encoding	<input type="radio"/> ON
<input type="checkbox"/> A holiday_Lastweekend Dummy encoding	<input type="radio"/> ON
<input type="checkbox"/> # Customer Life Avg-std rescaling	<input type="radio"/> ON
<input type="checkbox"/> # Frequency Avg-std rescaling	<input type="radio"/> ON
<input type="checkbox"/> # Mean Time Between Purchase Avg-std rescaling	<input type="radio"/> ON
<input type="checkbox"/> # Recency Avg-std rescaling	<input type="radio"/> ON
<input type="checkbox"/> # DateDiff Avg-std rescaling	<input type="radio"/> ON
<input type="checkbox"/> # Mean Time Between Invoice Avg-std rescaling	<input type="radio"/> ON
<input type="checkbox"/> # Avg Quantity Avg-std rescaling	<input type="radio"/> ON
<input type="checkbox"/> # Total Quantity Avg-std rescaling	<input type="radio"/> ON
<input type="checkbox"/> # Avg Spend Avg-std rescaling	<input type="radio"/> ON
<input type="checkbox"/> # Total Spend Avg-std rescaling	<input type="radio"/> ON
<input type="checkbox"/> # Total Spend stddev Avg-std rescaling	<input type="radio"/> ON
<input type="checkbox"/> # Total Spend first Avg-std rescaling	<input type="radio"/> ON
<input type="checkbox"/> # Total Spend last Avg-std rescaling	<input type="radio"/> ON
<input type="checkbox"/> # Changing Spend Avg-std rescaling	<input type="radio"/> ON
<input type="checkbox"/> # Target Spend Target variable	<input type="radio"/> @

Algorithms

Algorithms	CHANGE ALGORITHM PRESETS ▾	COPY TO...
Random Forest	<input checked="" type="checkbox"/> ON	
Gradient tree boosting	<input checked="" type="checkbox"/> ON	
Ordinary Least Squares	<input checked="" type="checkbox"/> ON	
Ridge Regression	<input checked="" type="checkbox"/> ON	
Lasso Regression	<input checked="" type="checkbox"/> ON	
LightGBM	<input checked="" type="checkbox"/> ON	
XGBoost	<input checked="" type="checkbox"/> ON	
Decision Tree	<input type="checkbox"/> OFF	
Support Vector Machine	<input type="checkbox"/> OFF	
Stochastic Gradient Descent	<input type="checkbox"/> OFF	
KNN	<input type="checkbox"/> OFF	
Extra Random Trees	<input checked="" type="checkbox"/> ON	
Neural Network	<input type="checkbox"/> OFF	
Lasso Path	<input type="checkbox"/> OFF	



SESSION 2			
<input checked="" type="checkbox"/>	Random forest (s2)	0.840	★
<input type="checkbox"/>	Gradient Boosted Trees (s2)	0.763	☆
<input type="checkbox"/>	Ordinary Least Squares (s2)	0.515	☆
<input type="checkbox"/>	Ridge (L2) regression (s2)	0.492	☆
<input type="checkbox"/>	Lasso (L1) regression (s2)	-0.002	☆
<input type="checkbox"/>	LightGBM (s2)	0.756	☆
<input type="checkbox"/>	XGBoost (s2)	0.482	☆
<input type="checkbox"/>	Extra trees (s2)	0.784	☆

เลือก Random Forest Model ในการ
คำนวณเพื่อว่ามีค่า R square สูงที่สุด

Result

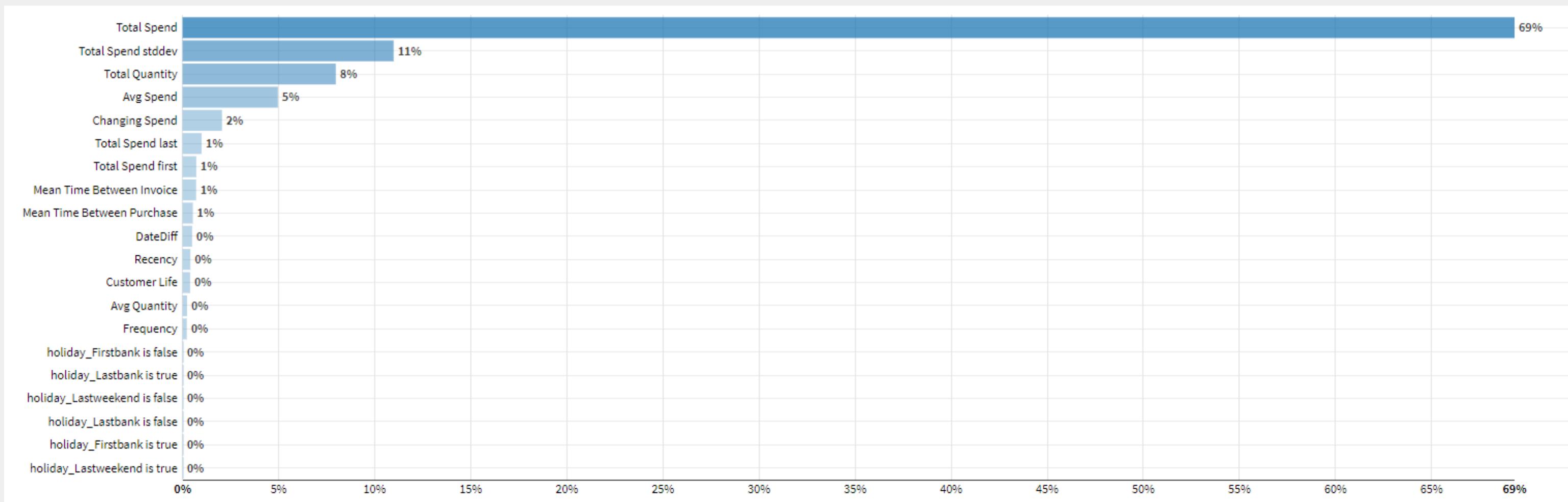
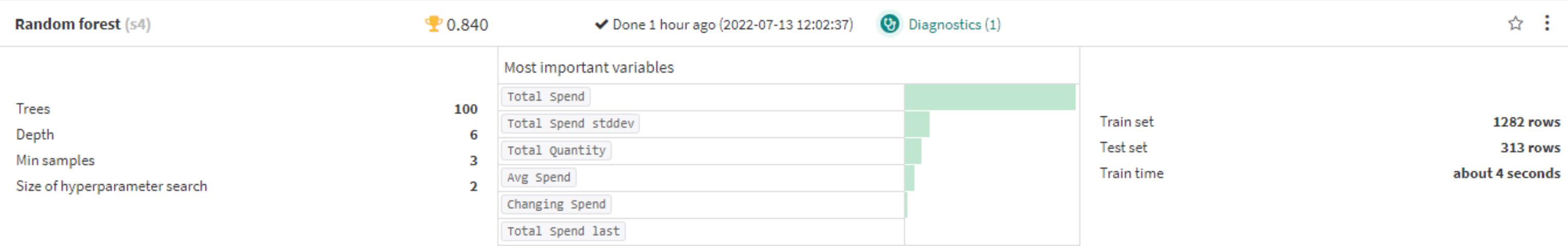
Random forest (s4)

0.840

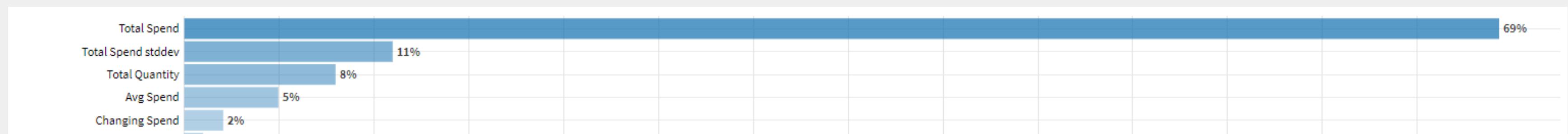
Done 1 hour ago (2022-07-13 12:02:37)

Diagnostics (1)

☆ ⋮



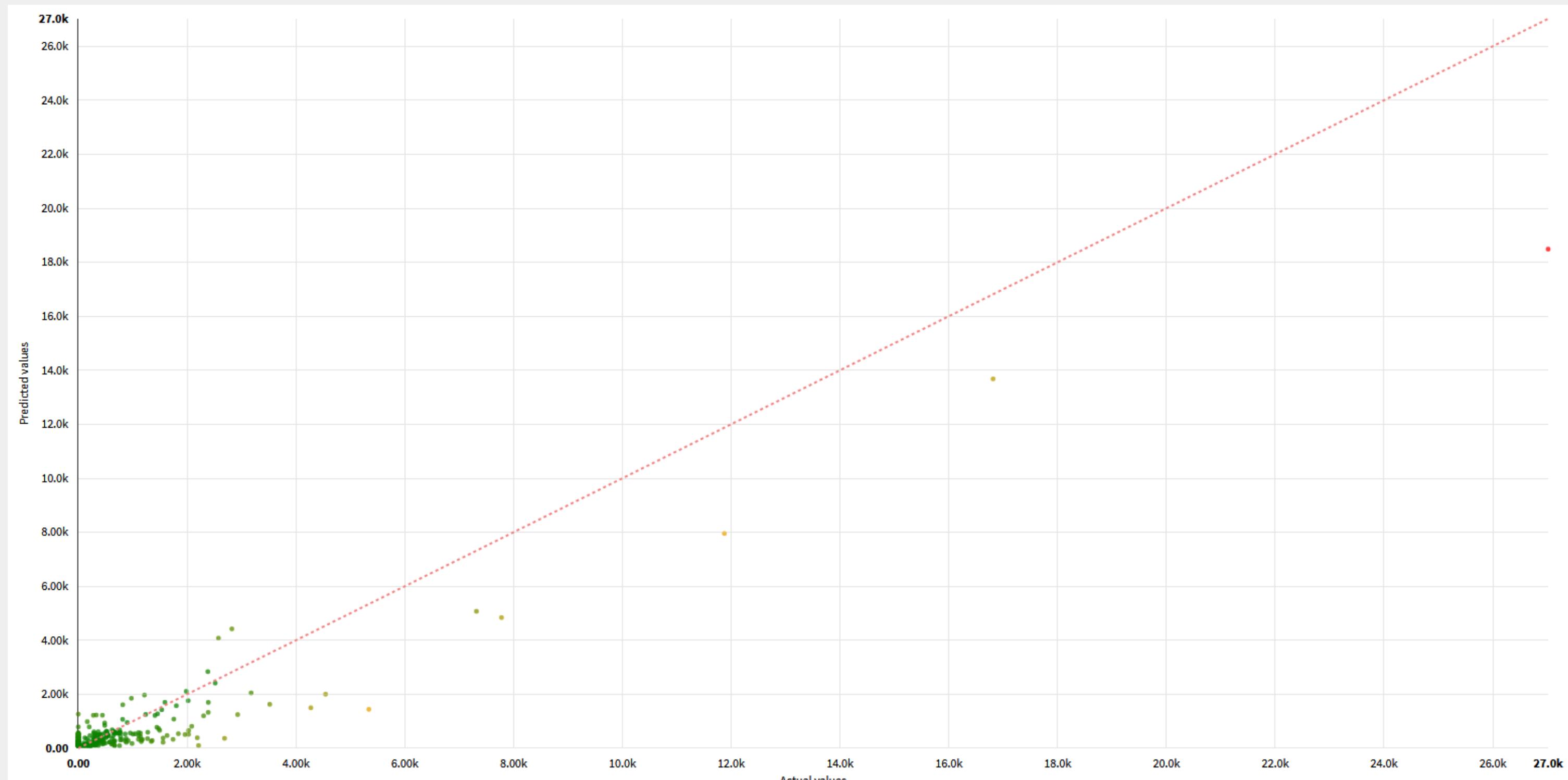
Variable Important



5 ตัวแปรที่ส่งผลกับความแม่นยำในการคำนวณของ Model มากที่สุดคือ

1. Total Spend เป็นตัวแปรที่บอกรถึงปริมาณการซื้อในเดือนที่ 1-3
2. Total Spend Stddev. ตัวแปรนี้บ่งบอกถึงความผันผวนในการซื้อของลูกค้า
3. Total Quantity บ่งบอกจำนวนสินค้าที่ซื้อทั้งหมดในเดือนที่ 1-3
4. Avg Spend บอกยอดซื้อเฉลี่ยในแต่ละครั้งของลูกค้า
5. Changing Spend คือสัดส่วนที่เกิดจากการนำยอดซื้อมาสูดหารด้วยครั้งแรก เพื่อดูว่ามีอัตราส่วนการเปลี่ยนแปลงอยู่ที่เท่า

Scatter Plot



Error Distribution

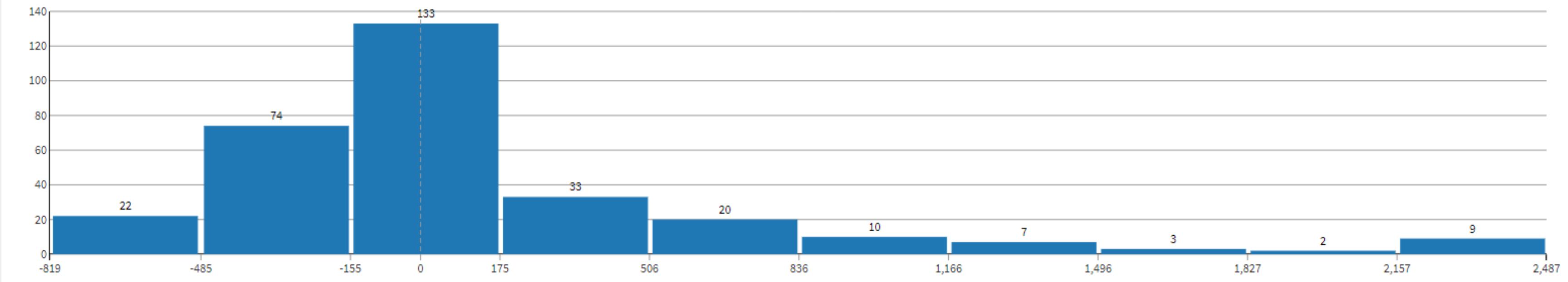
การกระจายตัวของค่า Error ส่วนใหญ่อยู่รอบๆ 0 ซึ่งจัดอยู่ในเกณฑ์ดี

Min. (raw)	Min.	25 th perc.	Median	75 th perc.	90 th perc.	Max.	Max. (raw)
-1600.2	-815.31	-203.40	-92.907	191.75	826.12	2487.2	8530.1
Average		110.49		Standard deviation		617.10	

Description

The errors (difference between predicted and actual values) should be centered around zero, and the distribution should be "narrow", i.e the spread of the error should be limited. More generally, the errors should be "normally" distributed around zero (the curve should look like a bell).

To reduce the effect of possible spurious outliers, error distribution is winsorized (clipped) at the 2nd and 98th percentiles.



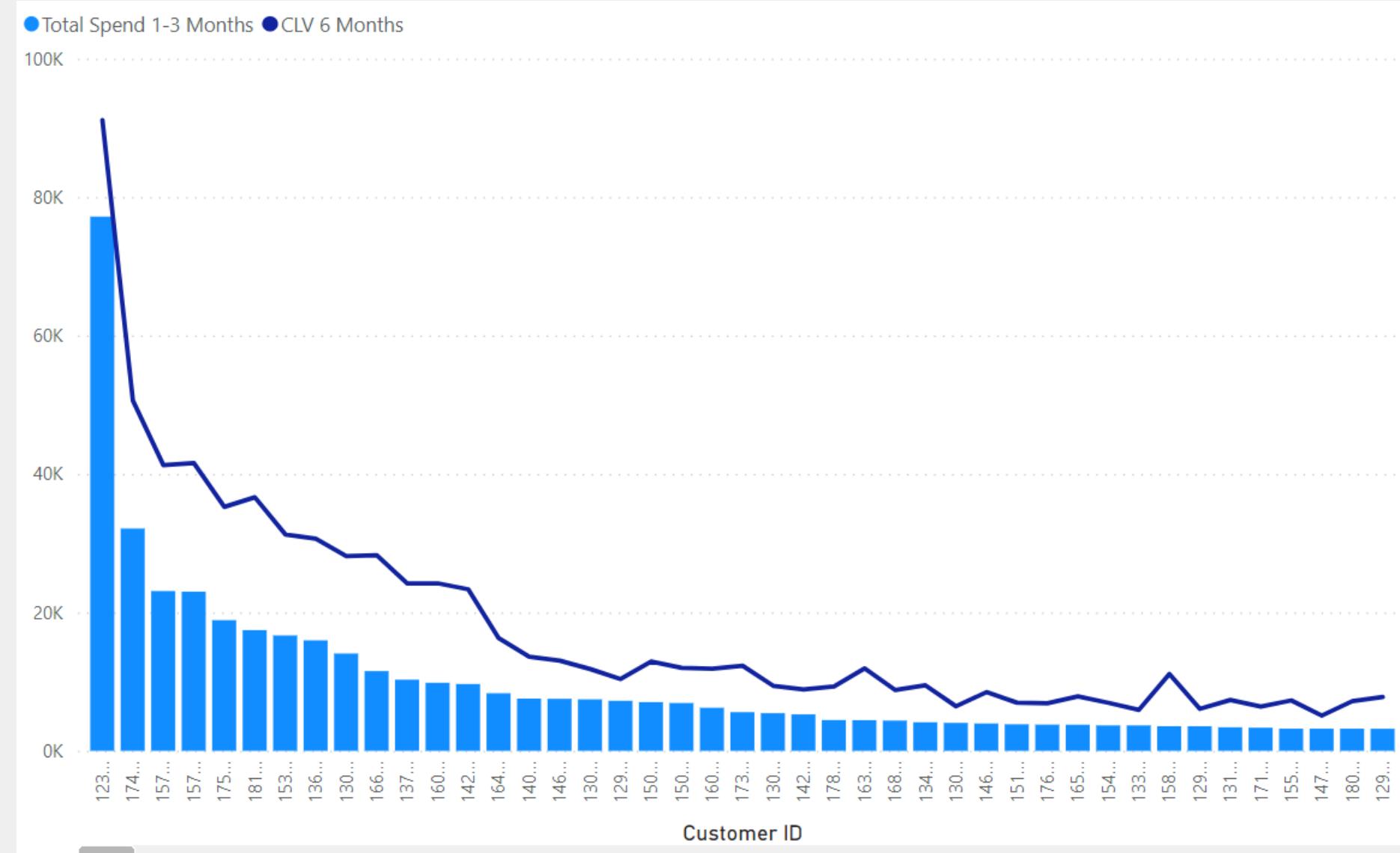


ส่วนที่ 4

Business Impact & Strategy



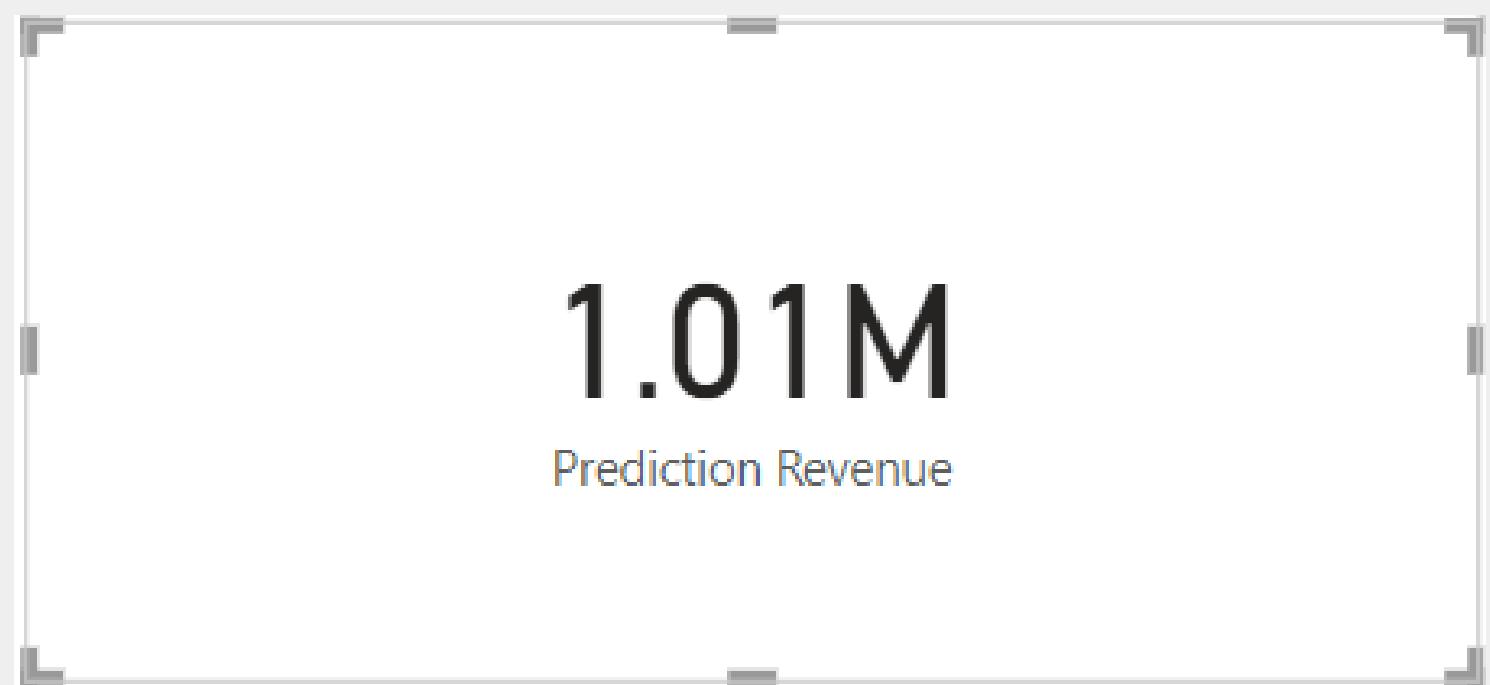
Business Strategy



Bar and Line Chart แสดงสัดส่วน
ระหว่าง CLV กึ่ง 6 เดือน กับยอด Total
Spend ใน 3 เดือนแรกในแต่ละ Customer

กลยุทธ์ที่ได้คือ เมื่อเวลาผ่านไป 1-2
เดือนแล้ว แต่ยอดซื้อของลูกค้าคงไหนยัง
ห่างไกลจากเส้น CLV 6 Months ที่เป็นเป้า
หมายแล้ว ก็จะให้ทีม Marketing ทำการ
Remind Customer ด้วยแคมเปญต่างๆ
หรืออาจจะใช้การให้ส่วนลดต่างๆ เพื่อกระตุ้น
ให้เกิดยอดซื้อ เป็นต้น

Business Impact



Thank You

Puphat Maneechan 6410414006

Management Analytics and Data Technologies