การทำนายลักษณะของลูกค้าที่ยกเลิกการใช้บริการ

Customer Churn Prediction

โดย

นางสาวณัฐกฤตา ป่งแก้ว

เลขทะเบียน 6009610947 สาขาสถิติ

เสนอ

ผู้ช่วยศาสตราจารย์ ดร. ประภาพร รัตนธำรง

รายงานนี้เป็นส่วนหนึ่งของรายวิชาการจำลองคอมพิวเตอร์และเทคนิคการพยากรณ์สำหรับธุรกิจ

CS358 COMPUTER SIMULATION AND FORECASTING TECHNIQUES IN BUSINESS

คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์

ภาคเรียนที่ 2 ปีการศึกษา 2563

คำนำ

การศึกษาในครั้งนี้เป็นการศึกษาเกี่ยวกับการทำนายลักษณะของลูกค้าที่มีแนวโน้มจะยกเลิกการใช้บริการ ในการดำเนินงานจะดำเนินงานตามแผน CRISP-DM ซึ่งในแต่ละขั้นตอนนั้นจะดำเนินการบน Google Cloud Platform ผ่านการใช้ Cloud Storage, AI Platform, Bigquery และ Data Studio ตั้งแต่กระบวนการนำเข้าข้อมูล เก็บข้อมูล ประมวลผล รวมไปถึงทำรายงานเพื่อนำเสนอ

การศึกษาครั้งนี้สำเร็จไปได้ด้วยดีด้วยความกรุณาจากผู้ช่วยศาสตราจารย์ ดร. ประภาพร รัตนธำรง ที่ได้กรุณาให้ความรู้ คำปรึกษา คำแนะนำที่เป็นประโยชน์ในการศึกษา ตลอดจนช่วยแก้ไขปัญหาต่าง ๆ ที่เกิดขึ้นด้วยความเอาใจใส่และให้ความเมตตาตลอดมา ขอขอบพระคุณเป็นอย่างสูงมา ณ ที่นี้ด้วย

สารบัญ

หน้า

ที่มาและความสำคัญ 1

เป้าหมาย 2

กรอบแนวคิด 2

ชุดข้อมูลและรายละเอียดของชุดข้อมูล 3

สถาปัตยกรรมระบบเบื้องต้น 4

แผนการดำเนินงาน 5

Business understanding 6

Data Understanding 6

Data Preparation 12

Modeling & Evaluation 13

Deployment 22

สรุปผลการศึกษา อภิปรายผล และข้อเสนอแนะ 26

สารบัญรูปภาพ

รูปภาพที่ 1 การเรียกดูข้อมูลในตาราง 7
รูปภาพที่ 2 data type ของข้อมูลที่จะนำมาวิเคราะห์ 8
รูปภาพที่ 3 แสดงจำนวน missing value 8
รูปภาพที่ 4 จำแนกลักษณะของลูกค้า 9
รูปภาพที่ 5 summary ของกลุ่มข้อมูลที่เป็นตัวเลขเชิงพรรณนา 9
รูปภาพที่ 6 box plot ของตัวแปรเชิงปริมาณ 10
รูปภาพที่ 7 bar chat ของตัวแปรเชิงคุณภาพ 11
รูปภาพที่ 8 correlation 13
รูปภาพที่ 9 เรียงลำดับค่า correlation 14
รูปภาพที่ 10 ตรวจสอบความสัมพันธ์ตัวแปรอิสระกับตัวแปรตาม 14
รูปภาพที่ 11 รายงานแสดงผล 22

สารบัญตาราง

ตารางที่ 1 รายละเอียดเกี่ยวกับชุดข้อมูล 3
ตารางที่ 2 แสดงแผนการดำเนินงาน 5
ตารางที่ 3 Feature Engineering 19
ตารางที่ 4 Confusion Matrix 20
ตารางที่ 5 Evaluation of model 20
ตารางที่ 6 Variance inflation factor 20
ตารางที่ 7 intercept & coefficient 21

CUSTOMER CHURN

- **ที่มาและความสำคัญ**

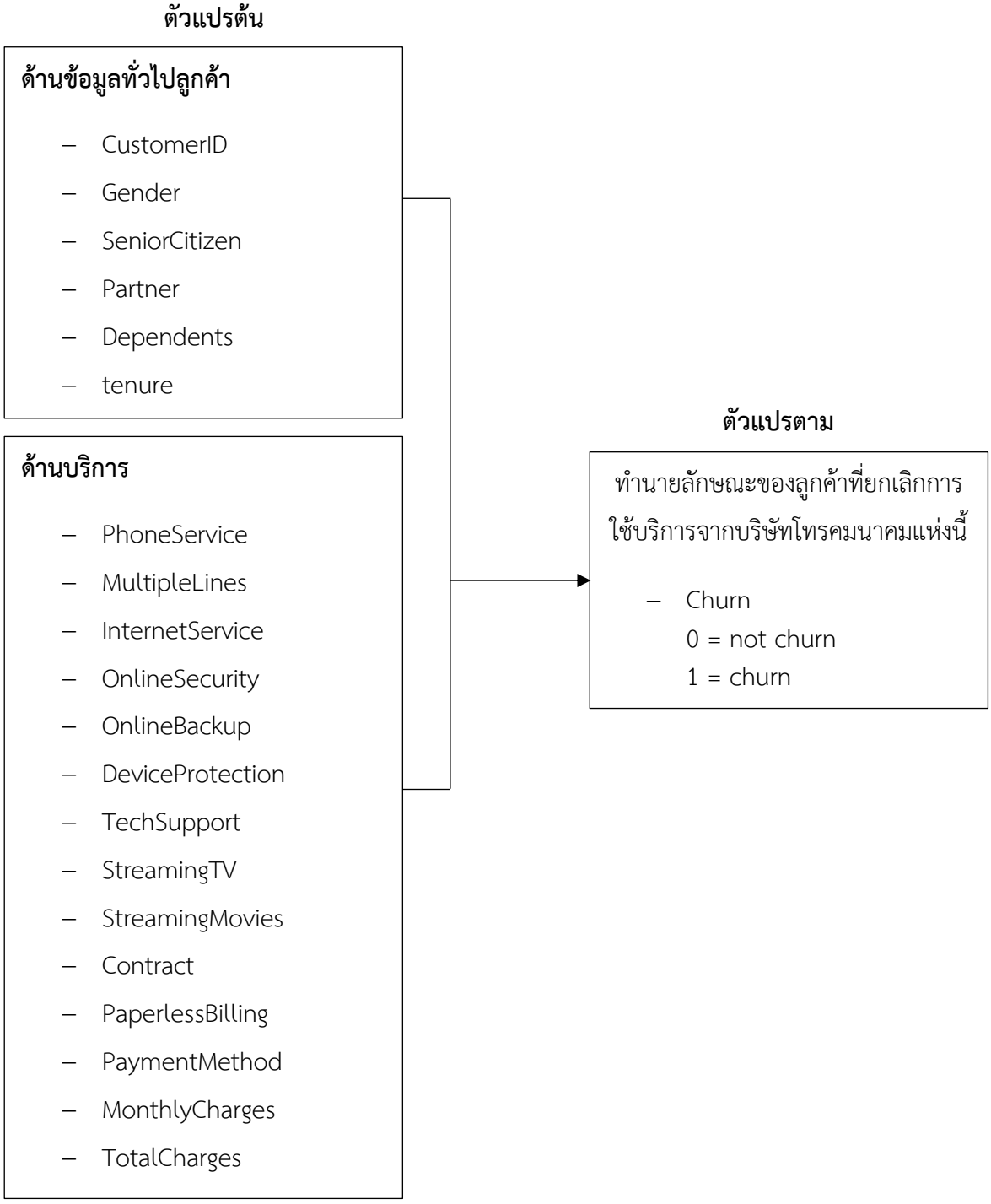
วงการธุรกิจในปัจจุบัน “Customer Churn” นับเป็นสิ่งที่นักธุรกิจต้องเผชิญและให้ความสำคัญเป็นอย่างมาก เนื่องจากไม่มีธุรกิจไหนที่จะเติบโตได้หากไม่มีลูกค้า เพราะรายได้หลักในแต่ละธุรกิจไม่ว่าจะธุรกิจใดล้วนมาจากลูกค้าทั้งนั้น ซึ่งการที่ลูกค้าเปลี่ยนใจยกเลิกการใช้บริการ (Customer churn) เป็นสิ่งที่ชี้ชัดให้เห็นว่าการดำเนินธุรกิจอาจกำลังมีปัญหา ถ้าหากปล่อยไว้โดยไม่หาสาเหตุหรือทำความเข้าใจในพฤติกรรมที่เปลี่ยนไปของลูกค้า อาจจะทำให้ธุรกิจเสียลูกค้าไปได้ง่าย ๆ จนส่งผลให้บริษัทต้องเสียค่าใช้จ่ายเพิ่มมากขึ้น เนื่องจากต้นทุนในการดึงดูดกลุ่มลูกค้าใหม่ตามหลัก CRM นั้นสูงกว่าการจัดการกับลูกค้าเก่ามาก จึงทำให้เห็นชัดว่าการรักษาลูกค้าเก่าให้อยู่กับบริษัทไปนาน ๆ เป็นสิ่งที่สมควรทำอย่างยิ่ง ดังนั้นในหลายธุรกิจจึงใช้ churn model เข้ามามีส่วนร่วมในการพัฒนา เพื่อวิเคราะห์และทำนายแนวโน้มที่ลูกค้าจะยกเลิกการใช้บริการ โดยเฉพาะอุตสาหกรรมด้านโทรคมนาคมที่ต้องเผชิญกับการแข่งขันอย่างมากจากผู้ให้บริการหลายราย ถ้าหากบริษัทสามารถคาดการณ์ได้ล่วงหน้าได้ว่าลูกค้ารายใดมีความเสี่ยงหรือมีแนวโน้มที่จะยกเลิกบริการก็จะเป็นประโยชน์อย่างมากกับบริษัท ซึ่งประโยชน์ของการวิเคราะห์และทำนายแนวโน้มนี้จะช่วยให้ธุรกิจเข้าใจลูกค้าได้ดีขึ้น ติดต่อกับลูกค้าที่มีความเสี่ยงได้โดยตรง และสามารถรับมือกับสถานการณ์เพื่อพยายามเปลี่ยนการตัดสินใจของลูกค้าที่จะยกเลิกการใช้บริการได้ เช่น การออก promotion เพื่อเพิ่มแรงจูงใจให้ลูกค้ากลับมาใช้บริการ หรือการมีข้อเสนอพิเศษเสนอไปยังลูกค้า ซึ่งจะสร้าง customer experience ที่ดีให้กับลูกค้าได้มาก

จากที่กล่าวไปข้างต้นทำให้ผู้จัดทำเห็นว่าการทำ churn model มีความสำคัญต่อวงการธุรกิจในปัจจุบันเป็นอย่างมาก ดังนั้นผู้จัดทำจึงอยากนำข้อมูลที่มีอยู่ให้เกิดประโยชน์สูงสุดโดยนำมาวิเคราะห์หาแนวโน้มที่จะทำให้ลูกค้ายังคงใช้บริการต่อไป เพื่อเป็นการลดความเสี่ยงจากการยกเลิกดังกล่าว ซึ่งข้อมูลที่จะนำมาศึกษาเป็นข้อมูลจากบริษัทโทรคมนาคมแห่งหนึ่ง ประกอบไปด้วยข้อมูลเกี่ยวกับลูกค้าที่ใช้บริการกว่า 7,000 ราย และเป้าหมายของการศึกษาครั้งนี้ คือ จะทำนายลักษณะของลูกค้าที่ยกเลิกการใช้บริการ โดยใช้เทคโนโลยี Big Data และ Machine Learning ที่ได้รับความนิยมอย่างมากในปัจจุบัน โดยเทคนิคนี้จะทำการตรวจหาลูกค้าที่มีแนวโน้มจะยกเลิกการใช้บริการ ซึ่งเหมาะสำหรับบริษัทที่ค้าขายด้วยระบบแพคเกจรายเดือนอย่างเช่นบริษัทโทรคมนาคมที่ให้บริการเกี่ยวกับโทรศัพท์และอินเทอร์เน็ตที่เราจะนำมาวิเคราะห์เป็นอย่างยิ่ง โดยจะนำข้อมูลลูกค้าในอดีตที่ผ่านมามาทำการศึกษา จำแนกข้อมูลตัวแปรที่สำคัญต่างๆและปรับปรุงคุณภาพของข้อมูลเพื่อนำเข้าแบบจำลอง โดยแบบจำลองที่ใช้ในการทำนายจะใช้ Classification หรือการแบ่งกลุ่ม ซึ่งการศึกษาครั้งนี้จะแบ่ง Target ออกเป็นสองกลุ่ม คือ กลุ่มลูกค้าปกติ และกลุ่มลูกค้าที่ยกเลิกการใช้บริการ โดยจะดำเนินการผ่าน Google Cloud Platform และผลที่ได้จากการทำนายจะถูกนำมาวิเคราะห์เพื่อทำให้เห็นกลยุทธ์ใหม่ ๆ ที่จะสามารถนำมาประยุกต์ใช้กับธุรกิจประเภทอุตสาหกรรมด้านโทรคมนาคมได้

- เป้าหมาย

ศึกษาปัจจัยที่เกี่ยวข้องและสร้างแบบจำลองเพื่อทำนายลักษณะของลูกค้าว่าลูกค้ารายใดมีความเสี่ยงหรือมีแนวโน้มที่จะยกเลิกบริการจากบริษัทโทรคมนาคมแห่งนี้

- กรอบแนวคิด



- ชุดข้อมูลและรายละเอียดของชุดข้อมูลที่จะใช้

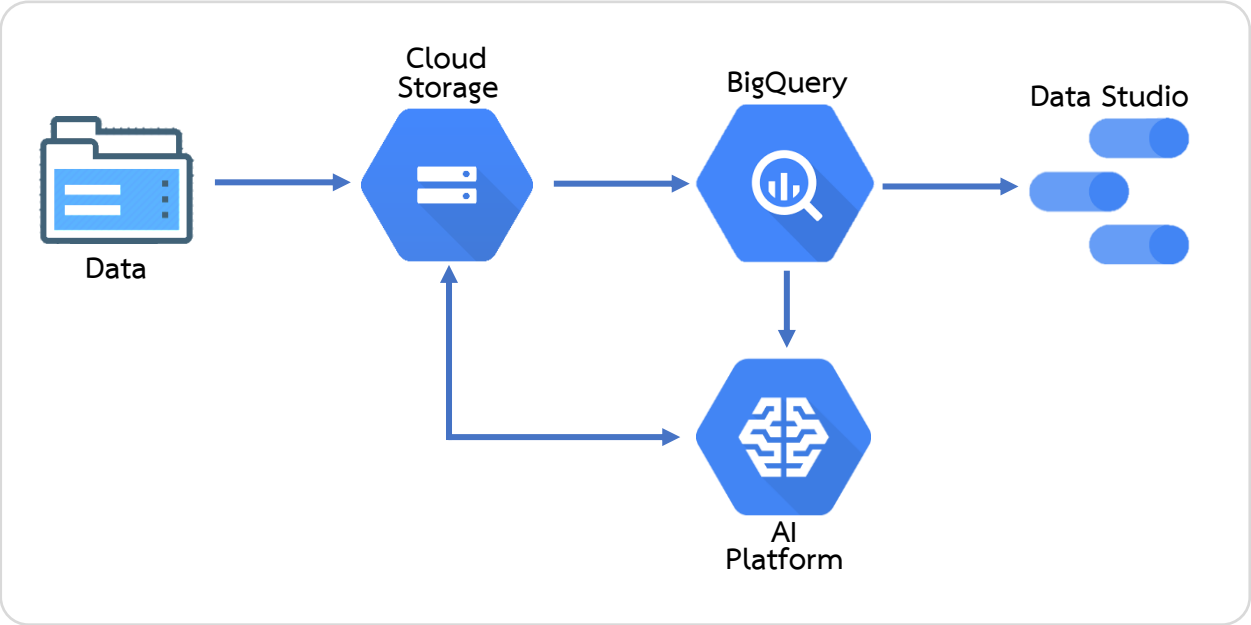
ชุดข้อมูลที่ใช้ในการศึกษาครั้งนี้ได้มาจาก Kaggle เป็นชุดข้อมูลเกี่ยวกับ Telco customer churn มีทั้งหมด 7043 case ซึ่งในชุดข้อมูลประกอบด้วย 21 คอลัมน์ มีรายละเอียดดังตารางที่ 1

ตารางที่ 1 รายละเอียดเกี่ยวกับชุดข้อมูล Telco customer churn

	ชื่อ	ชนิด	คำอธิบาย	ตัวอย่าง
1.	customerID	String	รหัสลูกค้า	7590-VHVEG
2.	gender	String	เพศ (Male, Female)	Female
3.	SeniorCitizen	int	ลูกค้าเป็นผู้สูงอายุที่มีอายุมากกว่า 65 ปีหรือไม่ (0=ไม่เป็น ,1= เป็น)	0
4.	Partner	String	ลูกค้าแต่งงานหรือไม่ (Yes, No)	Yes
5.	Dependents	String	ลูกค้ามีผู้ที่อยู่ในอุปการะหรือไม่ (Yes, No)	No
6.	tenure	int	จำนวนระยะเวลาที่ลูกค้าอยู่กับบริษัท	1
7.	PhoneService	String	ลูกค้ามีบริการโทรศัพท์หรือไม่ (Yes, No)	No
8.	MultipleLines	String	ลูกค้ามีบริการโทรศัพท์หลายสายหรือไม่ (Yes, No, No phone Service)	No phone Service
9.	InternetService	String	ลูกค้ามีบริการอินเทอร์เน็ตกับบริษัทหรือไม่ และเป็นโครงข่ายประเภทใด (Fiber, DSL ,No)	DSL
10.	OnlineSecurity	String	ลูกค้ามีบริการความปลอดภัยออนไลน์หรือไม่ (Yes, No, No internet service)	No
11.	OnlineBackup	String	ลูกค้ามีการสำรองข้อมูลออนไลน์หรือไม่ (Yes, No, No internet service)	Yes
12.	DeviceProtection	String	ลูกค้ามีแผนคุ้มครองอุปกรณ์หรือไม่ (Yes, No, No internet service)	No
13.	TechSupport	String	ลูกค้ามีทีมสนับสนุนทางเทคนิคหรือไม่ (Yes, No, No internet service)	No
14.	StreamingTV	String	ลูกค้ามีทีวีสตรีมมิ่งหรือไม่ (Yes, No, No internet service)	No
15.	StreamingMovies	String	ลูกค้ามีการสตรีมภาพยนตร์หรือไม่ (Yes, No, No internet service)	No
16.	Contract	String	ระยะสัญญาของลูกค้า (Month-to-month, One year, Two year)	Month-to-month

17.	PaperlessBilling	String	ลูกค้ามีการลดการใช้กระดาษในการเรียกเก็บค่าบริการหรือไม่ (Yes, No)	Yes
18.	PaymentMethod	String	วิธีการชำระเงินของลูกค้า (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))	Electronic check
19.	MonthlyCharges	Float	จำนวนเงินที่เรียกเก็บจากลูกค้ารายเดือน	29.85
20.	TotalCharges	Float	จำนวนเงินทั้งหมดที่เรียกเก็บจากลูกค้า	29.85
21.	Churn	String	ลูกค้าตัดสินใจยกเลิกหรือไม่ (Yes, No)	No

- สถาปัตยกรรมระบบ



1. นำข้อมูลที่จะทำการวิเคราะห์ไปเก็บไว้ที่ Cloud Storage โดยจะเก็บไว้ในรูปแบบไฟล์ .csv
   - git clone https://github.com/nesshipk/customer_churn
     gsutil -m cp *.csv gs://churn-project
2. ใช้ AI Notebook เพื่อทำ data understanding และ data preparation โดยมีการใช้ Cloud Storage และ Bigquery ร่วมด้วย หลังจากนั้นส่งข้อมูลกลับไปที่ Cloud Storage ในรูปแบบไฟล์ .csv เหมือนเดิม
3. นำเข้าข้อมูลที่ทำการ preparation เรียบร้อยแล้วมาบน AI Platform เพื่อทำการสร้างโมเดล เนื่องจากมี environment สำหรับทำโมเดล ไว้ค่อนข้างมาก หลังจากนั้นทำการวิเคราะห์จนได้โมเดลที่มีความถูกต้องแม่นยำ
4. ใช้ Data Studio ซึ่งจะเชื่อมต่อกับ BigQuery นำไปทำ visualize, report และ graph ต่าง ๆ

- แผนการดำเนินงาน

ตารางที่ 2 แสดงแผนการดำเนินงาน

	ขั้นตอนการดำเนินงาน	เมษายน				พฤษภาคม				มิถุนายน	
		1	2	3	4	1	2	3	4	1	2
1.	**หาข้อมูลที่สนใจ**										
2.	**Business Understanding** - เข้าใจปัญหาและแปลงปัญหาที่ได้ให้เป็น goal ที่จะนำมาวิเคราะห์										
3.	**Data Understanding** - รวบรวม ทำการตรวจสอบความถูกต้องของข้อมูลและสำรวจข้อมูลเบื้องต้น										
4.	**นำเสนอเค้าโครงโปรเจค**										
5.	**ปรับปรุงแก้ไข**										
6.	**Data Preparation** - data cleaning ต้องไม่มี missing value และทำ data transformation										
7.	**Modeling** - แบ่งข้อมูลเป็น 2 ส่วน คือ train data และ test data และนำ train data ไปสร้างแบบจำลองซึ่งใช้การจำแนกประเภทข้อมูล classification โดยใช้เทคนิค Logistic Regression										
8.	**รายงานความก้าวหน้า**										
9.	**ปรับปรุงแก้ไข**										
10.	**วิเคราะห์ผลที่ได้จาก model**										
11.	**Evaluation** - วัดประสิทธิภาพของโมเดล โดยใช้ confusion matrix										
12.	**Deployment** - สร้าง report เพื่อให้เข้าใจได้ง่าย										
13.	**นำเสนอผลงาน**										

กระบวนการในการดำเนินงานโดยใช้หลัก CRISP-DM

1. Business understanding

บริษัท Telco ในรัฐแคลิฟอร์เนีย ประเทศสหรัฐอเมริกา เป็นบริษัทที่ให้การบริการเกี่ยวกับโทรศัพท์และอินเทอร์เน็ตแก่ลูกค้า ซึ่งสิ่งที่สำคัญในการดำเนินธุรกิจนี้ คือ การหลีกเลี่ยงการยกเลิกสัญญาหรือการยกเลิกการใช้บริการของลูกค้า เนื่องจากในอุตสาหกรรมด้านโทรคมนาคมต้องเผชิญกับการแข่งขันที่รุนแรงจากผู้ให้บริการหลายราย ซึ่งต่างก็พยายามแข่งขันกันเพื่อให้สามารถรักษาฐานลูกค้าของตนเองเอาไว้ได้ ไม่ว่าจะเป็นการแข่งขันด้านราคา หรือยื่นข้อเสนอพิเศษต่าง ๆ ให้กับลูกค้า ถ้าหากบริษัทสามารถคาดการณ์ล่วงหน้าได้ว่าลูกค้ารายใดมีความเสี่ยงหรือมีแนวโน้มที่จะยกเลิกการใช้บริการก็จะเป็นประโยชน์อย่างมากกับบริษัท จะช่วยให้ธุรกิจเข้าใจลูกค้าได้ดีขึ้น ติดต่อกับลูกค้าที่มีความเสี่ยงได้โดยตรง และสามารถรับมือกับสถานการณ์เพื่อพยายามเปลี่ยนการตัดสินใจของลูกค้าที่จะยกเลิกการใช้บริการได้

2. Data Understanding

กลุ่มตัวอย่างที่ใช้ในการวิเคราะห์ครั้งนี้เป็นลูกค้าจำนวน 7,043 ราย ซึ่งข้อมูลชุดนี้ประกอบไปด้วย 20 features และ target variable (churn) แต่ละ features มีรายละเอียดดังนี้

- **customerID** ระบุรหัสเฉพาะที่ใช้ระบุลูกค้าแต่ละราย
- **gender** ระบุเพศของลูกค้า
- **SeniorCitizen** ระบุว่าลูกค้าเป็นผู้สูงอายุที่มีอายุมากกว่า 65 ปีหรือไม่
- **Partner** ระบุว่าลูกค้าแต่งงานแล้วหรือไม่
- **Dependent** ระบุว่าลูกค้ามีผู้ที่อยู่ในอุปการะหรือไม่ ซึ่งผู้ที่อยู่ในอุปการะจะหมายถึง เด็ก พ่อ แม่ ปู่ ย่า ตา ยาย
- **tenure** ระบุจำนวนเดือนทั้งหมดที่ลูกค้าอยู่กับบริษัท
- **PhoneService** ระบุว่าลูกค้าสมัครใช้บริการโทรศัพท์กับบริษัทหรือไม่
- **MultipleLines** ระบุว่าลูกค้าสมัครโทรศัพท์หลายสายกับบริษัทหรือไม่
- **InternetService** ระบุว่าลูกค้าสมัครใช้บริการอินเทอร์เน็ตกับบริษัทหรือไม่ ถ้าใช่เป็นอินเทอร์เน็ตที่ให้บริการผ่านโครงข่ายประเภทใด Fiber Optic หรือ DSL
- **OnlineSecurity** ระบุว่าลูกค้าสมัครใช้บริการรักษาความปลอดภัยออนไลน์เพิ่มเติมที่บริษัทได้ทำการจัดหาให้หรือไม่
- **OnlineBackup** ระบุว่าลูกค้าสมัครใช้บริการสำรองข้อมูลออนไลน์เพิ่มเติมที่บริษัทจัดหาให้หรือไม่
- **DeviceProtection** ระบุว่าลูกค้าสมัครแผนคุ้มครองอุปกรณ์เพิ่มเติมสำหรับอุปกรณ์อินเทอร์เน็ตที่บริษัทจัดหาให้หรือไม่

- **TechSupport** ระบุว่าลูกค้าสมัครแผนสนับสนุนด้านเทคนิคเพิ่มเติมจากบริษัทหรือไม่
- **StreamingTV** ระบุว่าลูกค้าใช้บริการอินเทอร์เน็ตเพื่อสตรีมรายการโทรทัศน์หรือไม่ ซึ่งบริษัทไม่เรียกเก็บค่าธรรมเนียมเพิ่มเติมสำหรับบริการนี้
- **StreamingMovies** ระบุว่าลูกค้าใช้บริการอินเทอร์เน็ตเพื่อสตรีมภาพยนตร์หรือไม่ ซึ่งบริษัทไม่เรียกเก็บค่าธรรมเนียมเพิ่มเติมสำหรับบริการนี้
- **Contract** ระบุประเภทสัญญาปัจจุบันของลูกค้าว่าเป็นประเภทเดือนต่อเดือน หนึ่งปี หรือสองปี
- **PaperlessBilling** ระบุว่าลูกค้าเลือกการเรียกเก็บค่าใช้บริการแบบลดการใช้กระดาษหรือไม่
- **PaymentMethod** ระบุวิธีการชำระเงินของลูกค้าว่าเป็นเช็คอิเล็กทรอนิกส์, เช็คทางไปรษณีย์, การโอนเงินระหว่างบัญชีธนาคาร (อัตโนมัติ), บัตรเครดิต (อัตโนมัติ)
- **MonthlyCharges** ระบุค่าบริการรายเดือนทั้งหมดของลูกค้าในปัจจุบัน
- **TotalCharges** ระบุค่าใช้จ่ายทั้งหมดของลูกค้า
- **Churn** ระบุว่าลูกค้ายกเลิกการใช้บริการจากบริษัทหรือไม่

ในการทำ Data Understanding เพื่อสำรวจข้อมูลเบื้องต้นก่อนที่จะนำไปใช้วิเคราะห์ต่อ จะใช้ AI Platform Notebooks (Jupyterlab) บน google cloud ซึ่งมีการเรียกใช้ Bigquery ที่ได้นำข้อมูลจาก Cloud storage มาสร้าง table ชื่อว่า data ใน dataset ที่ใช้ชื่อว่า churn และมีการใช้ Cloud storage โดยตั้งชื่อ bucket ว่า churn-project ซึ่งเป็นชื่อที่ไม่ซ้ำและเป็นชื่อที่ unique

ดูรายละเอียดเพิ่มเติมได้ที่ **data_understanding.ipynb**

```
%pip install google-cloud
%pip install google-cloud-storage
%pip install google-cloud-bigquery
%pip install pandas
from google.cloud import storage
from google.cloud import bigquery
import pandas as pd
```

ทำการเรียกดูข้อมูลในตารางเพื่อให้มั่นใจว่าข้อมูลทั้งหมดมี 7043 row และ 21 column ดังรูปภาพที่ 1

```
%%bigquery
SELECT *
FROM
    `churn.data`
```

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...	DeviceProtection	TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges	Churn
0	9732-OUYRN	Female	0	True	False	49	True	No	No	No internet service	...	No internet service	No internet service	No internet service	No internet service	One year	False	Credit card (automatic)	19.00	918.7	False
1	0661-KQHNK	Female	0	True	True	6	True	No	No	No internet service	...	No internet service	No internet service	No internet service	No internet service	One year	False	Credit card (automatic)	19.00	105.5	False
2	4709-LKHYG	Female	0	True	True	29	True	No	No	No internet service	...	No internet service	No internet service	No internet service	No internet service	One year	False	Electronic check	20.00	540.05	False
3	9824-QCJPK	Male	0	True	False	36	True	No	No	No internet service	...	No internet service	No internet service	No internet service	No internet service	One year	False	Mailed check	20.00	666.75	False
4	4716-MRVEN	Female	0	False	False	29	True	No	No	No internet service	...	No internet service	No internet service	No internet service	No internet service	One year	False	Mailed check	20.00	599.3	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
7038	5883-GTGVD	Male	0	False	False	19	True	Yes	Fiber optic	No	...	No	No	Yes	Yes	Month-to-month	True	Electronic check	99.95	1931.75	True
7039	5502-RLUYV	Female	0	True	True	69	True	Yes	Fiber optic	No	...	Yes	No	Yes	Yes	Month-to-month	True	Electronic check	103.95	7446.9	True
7040	3001-UNBTL	Male	1	True	True	29	True	Yes	Fiber optic	No	...	Yes	No	Yes	Yes	Month-to-month	True	Electronic check	103.95	2964.8	False
7041	5760-IFJOZ	Male	0	False	False	3	True	Yes	Fiber optic	No	...	Yes	No	Yes	Yes	Month-to-month	False	Mailed check	107.95	313.6	False
7042	2081-VEYEH	Male	0	False	False	3	True	No	Fiber optic	No	...	Yes	Yes	Yes	Yes	Month-to-month	True	Electronic check	107.95	318.6	False

7043 rows × 21 columns

รูปภาพที่ 1 การเรียกดูข้อมูลในตาราง

หลังจากนั้นทำการตรวจสอบ data type ของข้อมูลว่า data type ของข้อมูลแต่ละตัวแปรมีความถูกต้องหรือไม่ ซึ่งจากการตรวจสอบจะพบว่า TotalCharges มี data type ที่ไม่ถูกต้อง ดังรูปภาพที่ 2 เนื่องจากมี data type เป็น object ซึ่งในความจริงแล้วข้อมูลส่วนนี้ต้องมี data type เป็น float ดังนั้นจึงต้องทำการแปลงข้อมูลในส่วนนี้ก่อน เพื่อให้การวิเคราะห์มีประสิทธิภาพ

```
#check name of column, data type and number of rows
#load data from cloud storage to jupyter notebook
df = pd.read_csv('gs://churn-project/Telco-Customer-Churn.csv')
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 21 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   customerID        7043 non-null   object
 1   gender            7043 non-null   object
 2   SeniorCitizen     7043 non-null   int64
 3   Partner           7043 non-null   object
 4   Dependents        7043 non-null   object
 5   tenure            7043 non-null   int64
 6   PhoneService      7043 non-null   object
 7   MultipleLines     7043 non-null   object
 8   InternetService   7043 non-null   object
 9   OnlineSecurity    7043 non-null   object
 10  OnlineBackup      7043 non-null   object
 11  DeviceProtection  7043 non-null   object
 12  TechSupport       7043 non-null   object
 13  StreamingTV       7043 non-null   object
 14  StreamingMovies   7043 non-null   object
 15  Contract          7043 non-null   object
 16  PaperlessBilling  7043 non-null   object
 17  PaymentMethod     7043 non-null   object
 18  MonthlyCharges    7043 non-null   float64
 19  TotalCharges      7043 non-null   object
 20  Churn             7043 non-null   object
dtypes: float64(1), int64(2), object(18)
```

รูปภาพที่ 2 data type ของข้อมูลที่จะนำมาวิเคราะห์

```
#check missing value after that the data type for the "TotalCharges" column is float
df.isnull().sum()
```

```
customerID          0
gender              0
SeniorCitizen       0
Partner             0
Dependents          0
tenure              0
PhoneService        0
MultipleLines       0
InternetService     0
OnlineSecurity      0
OnlineBackup        0
DeviceProtection    0
TechSupport         0
StreamingTV         0
StreamingMovies     0
Contract            0
PaperlessBilling    0
PaymentMethod       0
MonthlyCharges      0
TotalCharges       11
Churn               0
```

รูปภาพที่ 3 แสดงจำนวน missing value

จะเห็นได้ว่า ถ้ามีการเปลี่ยนแปลง data type ของข้อมูลแต่ละตัวแปรให้ถูกต้อง จะพบว่า มีจำนวน missing value เกิดขึ้นในตัวแปร TotalCharges ทั้งสิ้น 11 แถว ดังรูปภาพที่ 3 ดังนั้นเพื่อให้การวิเคราะห์ข้อมูลเป็นไปได้อย่างถูกต้องแม่นยำ และมีประสิทธิภาพ จึงควร drop แถวที่มีค่า missing ออก ทำให้จำนวนข้อมูลจาก 7,043 แถว จะเหลือเพียง 7,032 แถวเท่านั้น

```
%%bigquery
SELECT
    Churn,count(Churn) AS numcustomer
FROM
    `churn.data_final`
GROUP BY Churn
```

จะพบว่าในจำนวนลูกค้าทั้ง 7,032 รายนั้น มีจำนวนลูกค้าที่ยกเลิกการใช้บริการกับบริษัท Telco เพียง 1,869 ราย ดังรูปภาพที่ 4 ซึ่งเมื่อเทียบกับจำนวนลูกค้าที่ยังคงใช้บริการต่อนับว่าเป็นจำนวนที่ไม่สมดุล เนื่องจากลูกค้าที่ใช้บริการต่อนั้นมีสูงถึง 5,163 คน การที่ข้อมูลอยู่ในลักษณะที่ไม่สมดุลแบบนี้ อาจจะทำให้การวิเคราะห์ค่อนข้างยาก เนื่องจากกลุ่มตัวอย่างที่เราสนใจ คือ ลูกค้ายกเลิกใช้บริการ (Churn=1) มีจำนวนที่ค่อนข้างน้อย

	Churn	numcustomer
0	0	5163
1	1	1869

รูปภาพที่ 4 จำแนกลักษณะของลูกค้า

จากการตรวจสอบค่าเฉลี่ย ส่วนเบี่ยงเบนมาตรฐาน ควอไทล์ ของตัวแปร tenure, MonthlyCharges และ TotalCharge จะเห็นได้ว่า ข้อมูลที่นำมาวิเคราะห์เป็นข้อมูลที่เหมาะสม ไม่มี Extreme Value เนื่องจากทั้งสามตัวแปรไม่มีค่าที่ติดลบ เพราะในทางปฏิบัติ tenure, MonthlyCharges และ TotalCharge ไม่สามารถเป็นค่าที่ติดลบได้ นอกจากนี้ จะเห็นได้ว่า TotalCharges มีค่าส่วนเบี่ยงเบนที่สูงมาก ดังรูปภาพที่ 5 ซึ่งทำให้เห็นว่าข้อมูลในตัวแปรนี้มีการกระจายมาก อย่างไรก็ตาม ในส่วนของตัวแปร SeniorCitizen จะเลือกที่จะไม่สนใจ เพราะถึงแม้ว่าตัวแปรจะมี data type เป็น int แต่ข้อมูลของตัวแปรนี้เป็นลักษณะข้อมูลเชิงคุณภาพ มีแค่ค่า 0 และ 1 เท่านั้น

	SeniorCitizen	tenure	MonthlyCharges	TotalCharges
count	7032.000000	7032.000000	7032.000000	7032.000000
mean	0.162400	32.421786	64.798208	2283.300441
std	0.368844	24.545260	30.085974	2266.771362
min	0.000000	1.000000	18.250000	18.800000
25%	0.000000	9.000000	35.587500	401.450000
50%	0.000000	29.000000	70.350000	1397.475000
75%	0.000000	55.000000	89.862500	3794.737500
max	1.000000	72.000000	118.750000	8684.800000

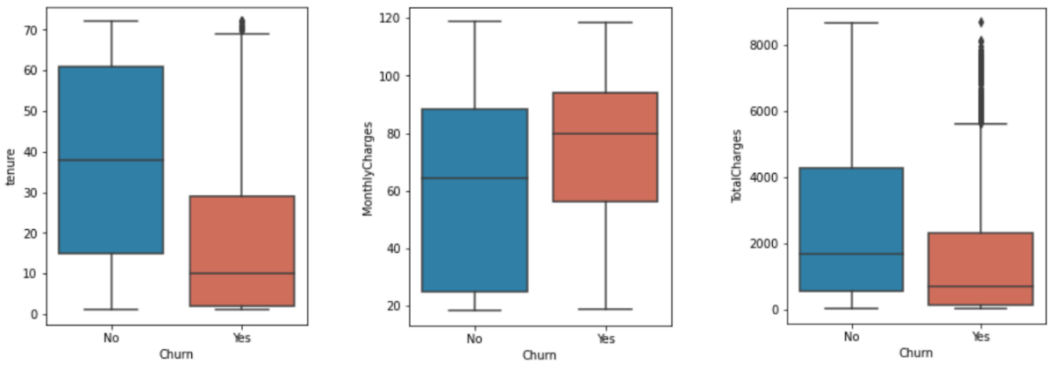
รูปภาพที่ 5 summary ของกลุ่มข้อมูลที่เป็นตัวเลขเชิงพรรณนา

สร้าง boxplot เปรียบเทียบระหว่างข้อมูล 2 กลุ่ม คือ กลุ่มลูกค้าที่ยังใช้บริการกับบริษัท Telco (not churn) และลูกค้าที่ยกเลิกการใช้บริการกับบริษัท Telco (churn) เพื่อใช้แสดงสาระที่สำคัญของข้อมูล ได้แก่ ค่ากลาง ค่าการกระจาย สัดส่วนของข้อมูลที่มากหรือน้อยกว่าค่ากลาง รวมทั้งข้อมูลที่อยู่ห่างจากกลุ่มมาก ๆ

```
import seaborn as sns
import matplotlib.pyplot as plt
cols = ["#1B86BA","#E36149"]
plt.figure(figsize = (4,5))
sns.boxplot(y=df['tenure'],x=df['Churn'],palette=cols)

plt.figure(figsize = (4,5))
sns.boxplot(y=df['MonthlyCharges'],x=df['Churn'],palette=cols)

plt.figure(figsize = (4,5))
sns.boxplot(y=df['TotalCharges'],x=df['Churn'],palette=cols)
```



รูปภาพที่ 6 box plot ของตัวแปรเชิงปริมาณ

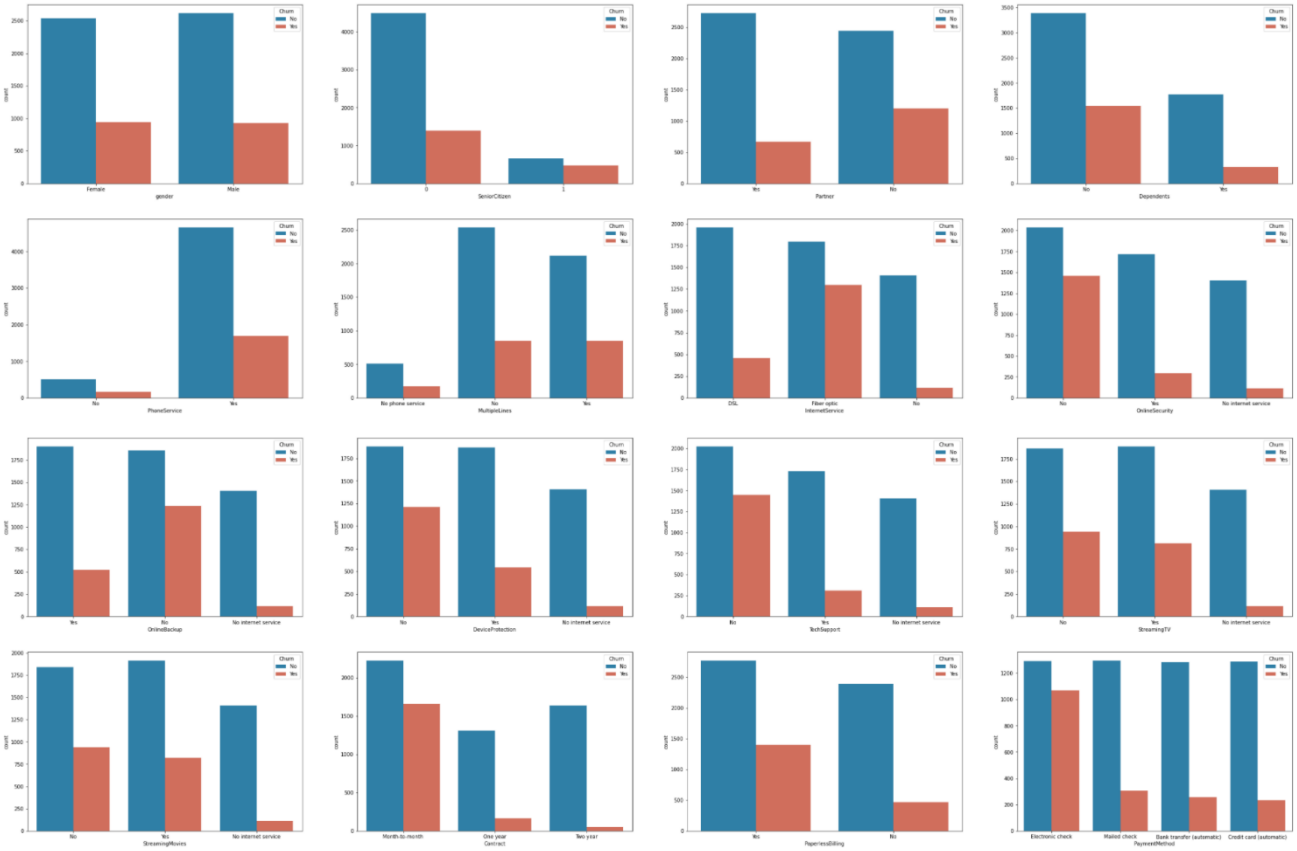
- **tenure** ลูกค้าที่ยกเลิกใช้บริการจะมีระยะเวลาเฉลี่ยที่อยู่กับบริษัทน้อยมากเพียงแค่ 10 เดือนเท่านั้น ในขณะที่ลูกค้าปกติมีระยะเวลาเฉลี่ยที่อยู่กับบริษัทประมาณ 40 เดือนและมีการกระจายของข้อมูลมาก จะเห็นได้ว่าลูกค้าที่ไม่ยกเลิกบริการจะมีระยะเวลาเฉลี่ยที่อยู่กับบริษัทค่อนข้างนานและทำให้เห็นว่าลูกค้าไม่ตัดสินใจที่จะยกเลิกบริการเมื่อระยะเวลาที่อยู่กับบริษัทมากกว่า 30 เดือน

- **MonthlyCharges** ลูกค้าที่ยกเลิกใช้บริการมีค่าใช้บริการรายเดือนสูงกว่ากลุ่มลูกค้าปกติ ซึ่งลูกค้าปกติจะมีค่าใช้บริการรายเดือนเฉลี่ยอยู่ราวๆ 60 USD ในขณะที่ลูกค้าที่ยกเลิกการใช้บริการมีค่าใช้บริการรายเดือนเฉลี่ยสูงถึง 80 USD และมีช่วงของ quartile ที่ต่ำกว่ามาก ซึ่งนั้นอาจจะเป็นอีกสาเหตุหนึ่งที่ทำให้ลูกค้ายกเลิกใช้บริการได้ เนื่องจากหากเปลี่ยนใจไปใช้บริการกับผู้ให้บริการเจ้าอื่นอาจจะทำให้ประหยัดได้มากกว่า

- **TotalCharges** จะเห็นว่ามีค่าเฉลี่ยที่ไม่แตกต่างกันมากเท่าไหร่ เป็นข้อมูลที่เปรียบเทียบค่อนข้างยาก เพราะเป็นข้อมูลการเรียกเก็บเงินทั้งหมด ซึ่งขึ้นอยู่กับระยะการใช้งานของลูกค้าแต่ละคน

หลังจากสร้าง box plot สำหรับตัวแปรเชิงปริมาณแล้ว ดังรูปภาพที่ 6 ต่อมาจะสร้าง box plot สำหรับตัวแปรเชิงคุณภาพ ได้แก่ ด้านข้อมูลทั่วไปของลูกค้า และด้านการบริการต่าง ๆ ดังรูปภาพที่ 7

```
plt.figure(figsize = (45,30))
cols = ["#1B86BA","#E36149"]
for i in enumerate(feature):
    plt.subplot(4,4,i[0]+1)
    sns.countplot(i[1],hue = 'Churn' , data = df,palette=cols)
```



รูปภาพที่ 7 bar chat ของตัวแปรเชิงคุณภาพ

ด้านข้อมูลทั่วไปของลูกค้า

- ไม่ว่าจะเพศหญิงหรือชาย จะมีจำนวนลูกค้าที่ยกเลิกใช้บริการที่ไม่แตกต่างกัน ดังนั้นเพศอาจจะไม่ใช่สาเหตุที่ทำให้ลูกค้ายกเลิกใช้บริการ

- ผู้สูงอายุจะมีอัตราในการยกเลิกใช้บริการที่สูงกว่า

- ลูกค้าที่แต่งงานแล้วและลูกค้าที่ไม่ได้อยู่ในอุปการะ จะมีจำนวนลูกค้าที่ยกเลิกใช้บริการที่สูงกว่า

ด้านการบริการต่างๆ

- Internet Service แบบ Fiber optic จะทำให้เกิดจำนวนที่ลูกค้ายกเลิกใช้บริการที่สูงกว่าแบบอื่นๆ

- การที่ลูกค้าไม่มีบริการเกี่ยวกับความปลอดภัยต่างๆ ไม่ว่าจะเป็น Online Security, Online Backup, Device Protection และ Technical Support จะทำให้โอกาสการยกเลิกใช้บริการสูงมาก

- ระยะเวลาของสัญญา (Contract) มีผลอย่างมากกับการที่ลูกค้าจะยกเลิกใช้บริการ ซึ่งยิ่งลูกค้ามีระยะสัญญาน้อยมากเท่าไหร่ โอกาสการยกเลิกใช้บริการย่อมมีมากกว่า

- วิธีการชำระเงินแบบ Electronic check จะมีจำนวนลูกค้าที่ยกเลิกใช้บริการที่สูงกว่าวิธีอื่นๆมาก โดยลูกค้าเกือบทั้งหมดตัดสินใจที่จะยกเลิกการใช้บริการ ทำให้เห็นว่าวิธีการชำระแบบนี้อาจจะมีข้อบกพร่องต่าง ๆ หรือมีปัญหาในระหว่างการทำธุรกรรมเกิดขึ้น

3. Data Preparation

ดำเนินการทำ data cleaning ก่อนการทำ modeling ซึ่งนับว่าเป็นกระบวนการที่สำคัญมาก หากทำการเตรียมข้อมูลได้ไม่ดี อาจส่งผลให้ผลการวิเคราะห์หรือการตีความจากการนำข้อมูลไปใช้ผิดเพี้ยนไปจากที่ควรจะเป็น

- เปลี่ยน data type ของข้อมูลให้ถูกต้อง โดยใช้ to_numeric ใน pandas

```
#see that the data type for the "TotalCharges" column is object
#so convert the values under the "TotalCharges" column into float
df['TotalCharges']=pd.to_numeric(df['TotalCharges'],errors='coerce')
```

- Drop missing value เพราะชุดข้อมูลที่มีข้อมูลที่สูญหาย (missing value) จะมีผลกระทบต่อโมเดลที่ถูกสร้างขึ้นมาเพื่อการทำนาย โดยใช้ dropna()

```
#see that the data in "TotalCharges" column has only 7032 from 7043 row
#so drop missing value
df = df.dropna()
```

- Convert Yes/No to 0,1 เพื่อแปลงข้อมูลให้เป็นตัวเลขจะได้สามารถสร้างโมเดลเพื่อการทำนายได้

```
#change the column of Yes/No to 1,0 and convert varlist into integer
varlist = ['Partner','Dependents','PhoneService','PaperlessBilling','Churn']
df[varlist] = df[varlist].replace(to_replace =['Yes','No'],value=['1','0'])
df[varlist] = df[varlist].astype(int)
df.head()
```

- Numeric encoding เพื่อแปลงข้อมูลให้เป็นตัวเลขจะได้สร้างโมเดลเพื่อการทำนายได้

```
#convert the column by using numerical encoding

#column 'gender'
df['gender'] = df['gender'].replace(to_replace =['Male','Female'],value=['1','0'])
df['gender'] = df['gender'].astype(int)
```

- One-hot แบบ dummy variable encoding โดยที่จำนวนคอลัมน์จะลดลงไป 1 column เนื่องจากข้อมูลเป็นแบบ categorical variable และข้อมูลไม่เรียงลำดับ ซึ่งการทำ one-hot จะทำให้โมเดล machine learning ทำงานได้ง่ายขึ้น

```
# one hot label
df = pd.get_dummies(df, columns = ['MultipleLines','InternetService','OnlineSecurity',
                                   'OnlineBackup','DeviceProtection','TechSupport',
                                   'StreamingTV','StreamingMovies','Contract',
                                   'PaymentMethod'], drop_first=True)
```

- Load data to cloud storage ไว้ใช้สำหรับขั้นตอนต่อไป คือ การสร้างโมเดลเพื่อการทำนาย

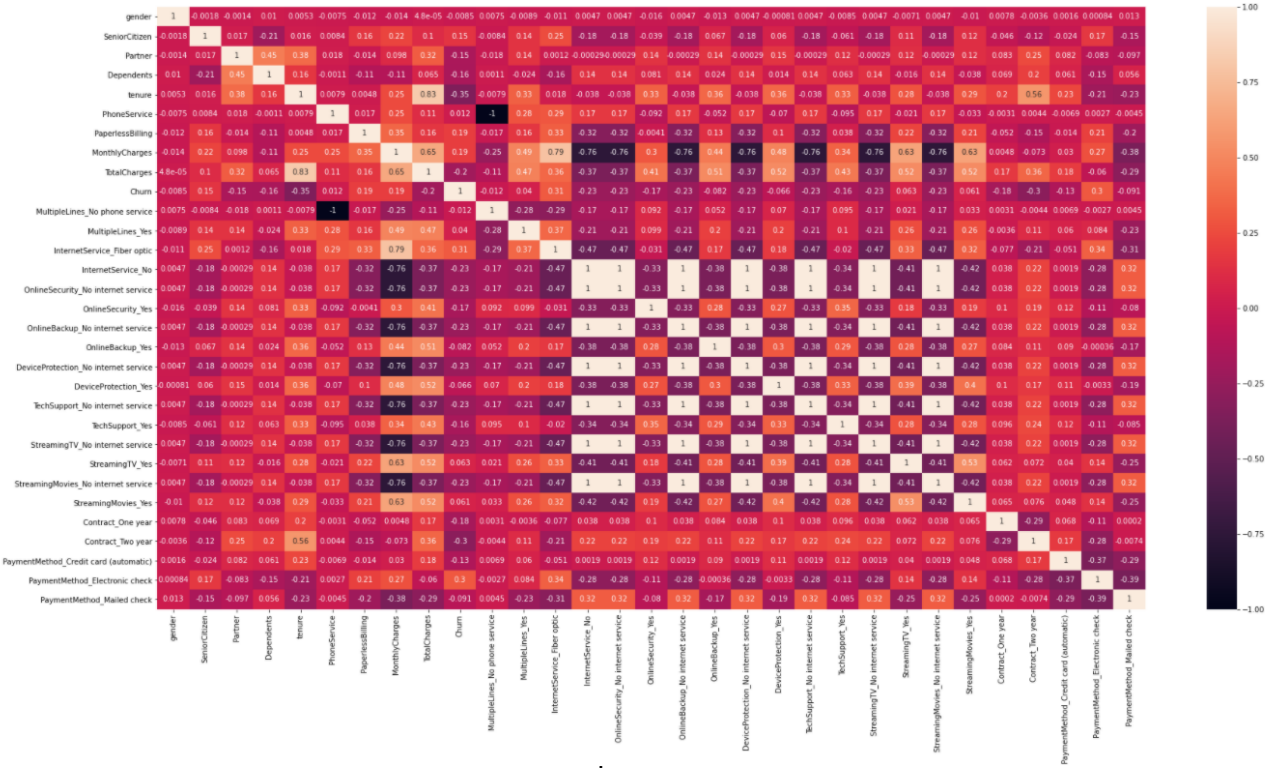
```
#load data to cloud storage in csv
df.to_csv('gs://churn-project/data_prep.csv')
```

ดูรายละเอียดเพิ่มเติมได้ที่ **data_preparation.ipynb**

4. Modeling & Evaluation

ในขั้นตอนแรก จะทำการดูค่าสัมประสิทธิ์สหสัมพันธ์ (Correlation) เพื่อดูความสัมพันธ์ระหว่างตัวแปรต้นและตัวแปรตาม

```
#create correlation plot
import seaborn as sns
import matplotlib.pyplot as plt
plt.figure(figsize = (30,15))
sns.heatmap(df.corr(),annot = True)
plt.show()
```

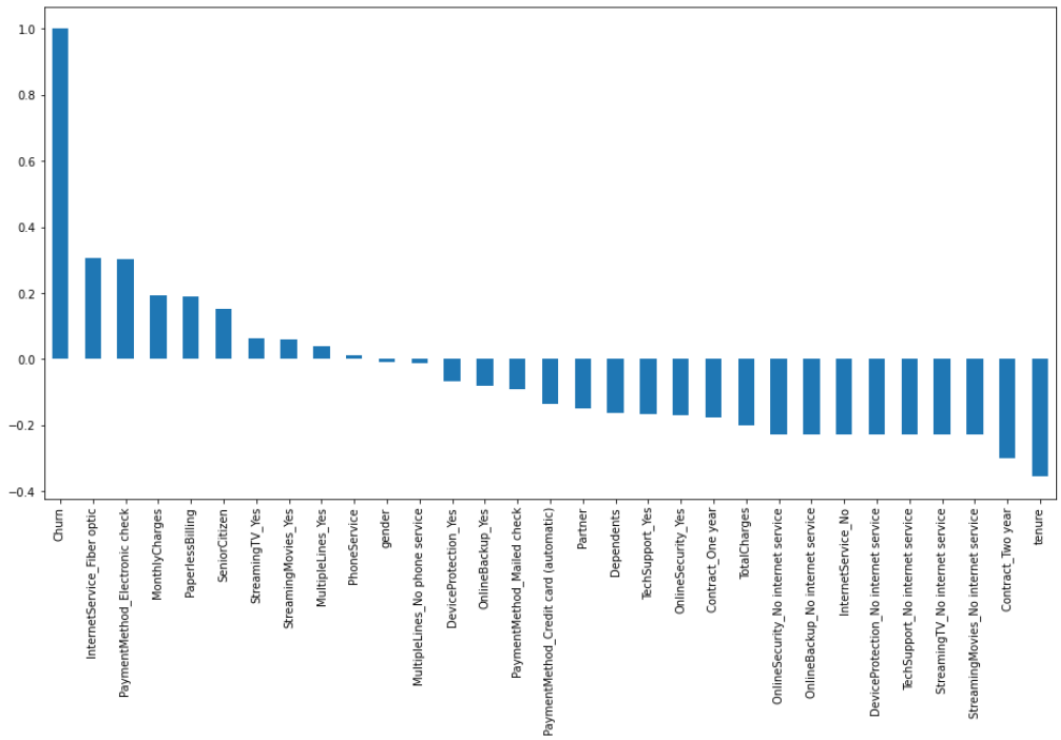


รูปภาพที่ 8 correlation

จากรูปภาพที่ 8 จะเห็นได้ว่า MonthlyCharges มีความสัมพันธ์ที่สูงมากกับตัวแปรอิสระอื่น ๆ ซึ่งในการสร้างโมเดลนั้น ตัวแปรอิสระไม่ควรจะมีความสัมพันธ์กันเอง หรือไม่เกิดปัญหา multicollinearity นั่นเอง รวมไปถึงไม่ควรเอา PhoneService เข้าโมเดลพร้อมกันกับ MultipleLine_No phone service เนื่องจากมีความสัมพันธ์กันสูง นอกจากนี้จะเห็นได้ว่า dummy variable ไม่มีตัวไหนมีความสัมพันธ์กันเองสูง ดังนั้นจึงไม่ต้องทำการ drop dummy variable

เพื่อทำการตรวจสอบว่าตัวแปรอิสระตัวไหนมีความสัมพันธ์กับตัวแปรตาม ดังนั้นจะทำการตรวจสอบที่ระดับนัยสำคัญ 0.05 โดยจะทำการตรวจสอบจากตัวแปรที่มีสหสัมพันธ์น้อยก่อน แสดงดังรูปภาพที่ 9

```
import matplotlib.pyplot as plt
%matplotlib inline
plt.figure(figsize=(16,8))
df.corr()['Churn'].sort_values(ascending = False).plot(kind='bar')
plt.show()
```



รูปภาพที่ 9 เรียงลำดับค่า correlation

ใช้ Statistical function (scipy.stats) เพื่อทดสอบสัมประสิทธิ์สหสัมพันธ์โดยเปรียบเทียบกับค่านัยสำคัญ 0.05 โดยผลลัพธ์ที่แสดงจะแสดงค่า correlation (ทางซ้าย) และค่า p-value (ทางขวา)

```
import scipy.stats
# 1. gender
corr = scipy.stats.pearsonr(df.gender,df.Churn)
corr
# p-value > 0.05
```

```
(-0.008544643224946243, 0.47373573732654467)
```

```
# 2. PhoneService
corr = scipy.stats.pearsonr(df.PhoneService,df.Churn)
corr
# p-value > 0.05
```

```
(0.011691398865421575, 0.32695528135874624)
```

```
# 3. MultipleLines_No phone service
corr = scipy.stats.pearsonr(df['MultipleLines_No phone service'],df.Churn)
corr
# p-value > 0.05
```

```
(-0.011691398865421568, 0.32695528135874624)
```

```
# 4. MultipleLines_Yes
corr = scipy.stats.pearsonr(df['MultipleLines_Yes'],df.Churn)
corr
# p-value < 0.05
```

```
(0.04003273987252132, 0.0007857240573236339)
```

รูปภาพที่ 10 ตรวจสอบความสัมพันธ์ตัวแปรอิสระกับตัวแปรตาม

จะเห็นได้ว่า gender , PhoneService และ MultipleLines_No phone service ไม่มีความสัมพันธ์กับตัวแปรตามที่ระดับนัยสำคัญ 0.05 ดังรูปภาพที่ 10 เนื่องจาก p-value มีค่ามากกว่าระดับนัยสำคัญ 0.05 ดังนั้น จึงไม่ควรนำตัวแปร gender , PhoneService และ MultipleLines_No phone service เข้าไปในการสร้างโมเดล

ในการสร้างโมเดลในการวิเคราะห์ครั้งนี้จะใช้ Logistic Regression เนื่องจาก Logistic Regression นิยมใช้กับปัญหา Binary Classification ซึ่งทำนาย target variable ที่มี 2 class และในการวิเคราะห์ครั้งนี้จะแบ่ง training data และ test data ออกเป็น 80 : 20 เนื่องจากข้อมูลตัวแปรตามไม่สมดุลกัน จึงอยากให้ในการ train model ใช้ข้อมูลจำนวนมากเพื่อให้สามารถวิเคราะห์ออกมาได้อย่างมีประสิทธิภาพมากกว่าแบ่งข้อมูลออกเป็น 70 : 30

```
#not bring customerID, gender, PhoneService and MultipleLines_No phone service into model
X = df.drop(['customerID','gender','PhoneService','MultipleLines_No phone service',
             'InternetService_No','OnlineSecurity_No internet service',
             'OnlineBackup_No internet service','DeviceProtection_No internet service',
             'TechSupport_No internet service','StreamingTV_No internet service',
             'StreamingMovies_No internet service','Churn'] ,axis='columns')
y = df['Churn']
X_train, X_test, y_train, y_test = train_test_split(X, y, train_size=0.8, test_size=0.2, random_state=20)
```

```
model = LogisticRegression(solver='lbfgs', max_iter=500)
model.fit(X_train, y_train)
model.score(X_train, y_train)
```

```
0.7982222222222223
```

```
metrics.confusion_matrix(y_test, model.predict(X_test))
```

```
array([[922, 120],
       [159, 206]])
```

```
metrics.accuracy_score(y_test, model.predict(X_test))
```

```
0.8017057569296375
```

```
from sklearn.metrics import confusion_matrix , classification_report
print(classification_report(y_test,model.predict(X_test)))
```

```
              precision    recall  f1-score   support

           0       0.85      0.88      0.87      1042
           1       0.63      0.56      0.60       365

    accuracy                           0.80      1407
   macro avg       0.74      0.72      0.73      1407
weighted avg       0.80      0.80      0.80      1407
```

not churn
churn

เมื่อทำการพิจารณาแยกทีละคลาสจะเห็นได้ว่า โมเดลนี้ตอบคำถามเกี่ยวกับการทำนายลักษณะของลูกค้าที่ยังใช้บริการกับบริษัท telco ได้ดีกว่าการทำนายลักษณะของลูกค้าที่ยกเลิกการใช้บริการ เนื่องจากค่า precision, recall และ f1-score ที่ค่า y เป็น 0 (not churn) มีค่าที่มากกว่าค่อนข้างมาก

```
from statsmodels.stats.outliers_influence import variance_inflation_factor
vif = pd.DataFrame()
vif['Features'] = X_train.columns
vif['VIF'] = [variance_inflation_factor(X_train.values, i) for i in range(len(X_train.columns))]
vif['VIF']
```

```
0      1.375866
1      2.833565
2      1.961732
3     15.872841
4      2.918609
5     38.515298
6     17.413823
7      2.671748
8      8.806129
9      2.162190
10     2.363046
11     2.440090
12     2.269508
13     3.158863
14     3.213663
15     1.929123
16     3.178950
17     1.853440
18     2.743710
19     1.872552
Name: VIF, dtype: float64
```

จะเห็นได้ว่าเมื่อตรวจสอบค่า VIF จะพบว่า MonthlyCharges มีค่า VIF สูงมากถึง 38.515298 ดังนั้นจึงควรทำการ drop MonthlyCharges ออกจากโมเดล เนื่องจากค่า VIF แสดงให้เห็นว่าถ้าตัวแปรทำนายนั้นมีความสัมพันธ์กันจะทำให้ความแปรปรวนของค่าสัมประสิทธิ์ของตัวแบบการถดถอยจะมีค่าเพิ่มขึ้น ซึ่งเมื่อความแปรปรวนเพิ่มขึ้นจึงหมายถึงว่าไม่ดี เพราะต้องการความแม่นยำในการประมาณค่า และเมื่อความแปรปรวนของเพิ่มขึ้นแปลว่าความน่าเชื่อถือของตัวแบบลดลง

```
#drop high vif
X_train_2 = X_train.drop(['MonthlyCharges'] ,axis='columns')
X_test_2 = X_test.drop(['MonthlyCharges'] ,axis='columns')
```

```
model.fit(X_train_2, y_train)
model.score(X_train_2, y_train)
```

```
0.8014222222222223
```

```
metrics.confusion_matrix(y_test, model.predict(X_test_2))
```

```
array([[923, 119],
       [159, 206]])
```

```
#final model 2
metrics.accuracy_score(y_test, model.predict(X_test_2))
```

```
0.8024164889836531
```

```
from sklearn.metrics import confusion_matrix , classification_report
print(classification_report(y_test,model.predict(X_test_2)))
```

```
              precision    recall  f1-score   support

           0       0.85      0.89      0.87      1042
           1       0.63      0.56      0.60       365

    accuracy                           0.80      1407
   macro avg       0.74      0.73      0.73      1407
weighted avg       0.80      0.80      0.80      1407
```

```
from statsmodels.stats.outliers_influence import variance_inflation_factor
vif = pd.DataFrame()
vif['Features'] = X_train_2.columns
vif['VIF'] = [variance_inflation_factor(X_train_2.values, i) for i in range(len(X_train_2.columns))]
vif['VIF']
```

```
0      1.372444
1      2.820074
2      1.947476
3     14.999978
4      2.710508
5     16.609214
6      2.356952
7      3.400177
8      1.812915
9      2.137647
10     2.241759
11     1.953781
12     2.713725
13     2.776221
14     1.928517
15     3.161289
16     1.640283
17     2.229818
18     1.414247
Name: VIF, dtype: float64
```

หลังจากทำการวิเคราะห์อีกครั้งโดยการ drop ตัวแปรที่มีค่า VIF สูง อย่าง MonthlyCharges ออกจากโมเดล ยังพบว่ายังมีตัวแปรที่มีค่า VIF สูงอยู่ นั่นก็คือ TotalCharges ซึ่งมีค่า VIF สูงมากถึง 16.609214 ดังนั้นจึงควรทำการ drop TotalCharges ออกจากโมเดล

```
#drop high vif
X_train_3 = X_train.drop(['MonthlyCharges','TotalCharges'] ,axis='columns')
X_test_3 = X_test.drop(['MonthlyCharges','TotalCharges'] ,axis='columns')
```

```
model.fit(X_train_3, y_train)
model.score(X_train_3, y_train)
```

```
0.8021333333333334
```

```
metrics.confusion_matrix(y_test, model.predict(X_test_3))
```

```
array([[922, 120],
       [161, 204]])
```

```
#final model 3
metrics.accuracy_score(y_test, model.predict(X_test_3))
```

```
0.8002842928216063
```

```
from sklearn.metrics import confusion_matrix , classification_report
print(classification_report(y_test,model.predict(X_test_3)))
```

```
              precision    recall  f1-score   support

           0       0.85      0.88      0.87      1042
           1       0.63      0.56      0.59       365

    accuracy                           0.80      1407
   macro avg       0.74      0.72      0.73      1407
weighted avg       0.79      0.80      0.80      1407
```

```
from statsmodels.stats.outliers_influence import variance_inflation_factor
vif = pd.DataFrame()
vif['Features'] = X_train_3.columns
vif['VIF'] = [variance_inflation_factor(X_train_3.values, i) for i in range(len(X_train_3.columns))]
vif['VIF']
```

```
0     1.368632
1     2.818457
2     1.937611
3     7.129092
4     2.655310
5     2.290586
6     2.813020
7     1.741986
8     2.004163
9     2.144156
10    1.886866
11    2.582016
12    2.634311
13    1.917666
14    3.094172
15    1.572094
16    1.992991
17    1.304793
Name: VIF, dtype: float64
```

หลังจากทำการวิเคราะห์อีกครั้งโดยการ drop ตัวแปรที่มีค่า VIF สูง อย่าง MonthlyCharges และ TotalCharges ออกจากโมเดล จะเห็นได้ว่าเมื่อตรวจสอบค่า VIF จะพบว่า tenure มีค่า VIF 7.129092 ซึ่งมีค่ามากกว่า 5 เนื่องจากโมเดลที่ดีนั้นควรมีค่า VIF ไม่เกิน 5 บางทฤษฎีกำหนดให้ไม่เกิน 10 แต่ในที่นี้จะใช้เกณฑ์ VIF ไม่เกิน 10 เนื่องจากได้ทำการตรวจสอบแล้วว่าถ้าหากทำการ drop tenure ออกจากโมเดล ทำให้เปอร์เซ็นต์การทำนายลูกค้าที่จะยกเลิกใช้บริการของบริษัท Telco นั้นลดลง ดังนั้น โมเดลในครั้งที่ 3 จึงมีความเหมาะสมแล้วในการทำนายลักษณะของลูกค้าของบริษัท telco

ต่อมาจะทำ Feature Engineering เพื่อให้โมเดลเรียนรู้ได้ดีขึ้น ซึ่งถือเป็นการเพิ่มประสิทธิภาพ Machine learning Model โดยในการวิเคราะห์ครั้งนี้จะใช้ค่า precision, recall, f1-score ของ class ที่ค่า y เป็น 1 และค่า accuracy เพื่อตัด Feature ที่ไม่เกี่ยวข้องทิ้งไป ซึ่งจะแสดงผลใน column สุดท้าย หากเห็นว่าควรตัด Feature นั้นออกจากโมเดลจะแสดงเครื่องหมาย - แต่ถ้าไม่ควรตัดจะแสดงเครื่องหมาย +

หลังจากการทำ Feature Engineering จะพบว่า ควรทำการเอา Feature เหล่านี้ ออกจากโมเดล เนื่องจาก เมื่อทำการเอา feature เหล่านี้ออก ทำให้โมเดลมีการเรียนรู้ที่ดีขึ้น สามารถทำนายได้อย่างถูกต้อง ดังตารางที่ 3

- Partner
- Dependents
- OnlineBackup_Yes
- DeviceProtection_Yes
- StreamingTV_Yes
- StreamingMovies_Yes
- Contract_Two year
- PaymentMethod_Credit card (automatic)
- PaymentMethod_Electronic check
- PaymentMethod_Mailed check

ตารางที่ 3 Feature Engineering

#	Variable	precision	recall	f1-score	accuracy	
1.	SeniorCitizen, Partner, Dependents, Tenure, PaperlessBilling, MultipleLines_Yes, InternetService_Fiber optic, OnlineSecurity_Yes, OnlineBackup_Yes, DeviceProtection_Yes, TechSupport_Yes, StreamingTV_Yes, StreamingMovies_Yes, Contract_One year, Contract_Two year, PaymentMethod_Credit card (automatic) PaymentMethod_Electronic check PaymentMethod_Mailed check	0.63	0.56	0.59	0.8003	N / A
2.	Remove SeniorCitizen	0.63	0.55	0.59	0.7989	+
3.	Remove Partner	0.64	0.56	0.60	0.8053	-
4.	Remove Dependents	0.64	0.57	0.60	0.8060	-
5.	Remove tenure	0.59	0.50	0.54	0.7810	+
6.	Remove PaperlessBilling	0.63	0.55	0.59	0.7989	+
7.	Remove MultipleLines_Yes	0.63	0.55	0.59	0.7996	+
8.	Remove InternetService_Fiber optic	0.62	0.52	0.56	0.7932	+
9.	Remove OnlineSecurity_Yes	0.62	0.55	0.59	0.7974	+
10.	Remove OnlineBackup_Yes	0.63	0.56	0.60	0.8017	-
11.	Remove DeviceProtection_Yes	0.63	0.56	0.60	0.8017	-
12.	Remove TechSupport_Yes	0.62	0.54	0.58	0.7953	+
13.	Remove StreamingTV_Yes	0.64	0.56	0.59	0.8024	-
14.	Remove StreamingMovies_Yes	0.63	0.57	0.60	0.8017	-
15.	Remove Contract_One year	0.63	0.53	0.58	0.7989	+
16.	Remove Contract_Two year	0.64	0.54	0.58	0.8003	-
17.	Remove PaymentMethod_Credit card (automatic)	0.63	0.57	0.60	0.8031	-
18.	Remove PaymentMethod_Electronic check	0.64	0.56	0.59	0.8024	-
19.	Remove PaymentMethod_Mailed check	0.63	0.56	0.59	0.8003	-

จากการนำ Feature ที่ไม่เกี่ยวข้องออกจากโมเดล ทำเห็นว่าโมเดลมีการเรียนรู้ที่ดีขึ้น สามารถทำนายได้ถูกต้องมากขึ้น แสดงดังตารางที่ 4

ตารางที่ 4 Confusion Matrix

Actual/Predict	not churn	churn
not churn	973	105
churn	172	193

จากตารางที่ 5 พบว่า โมเดลหลังจากการทำ Feature Engineering มี accuracy score เท่ากับ 0.8031 ซึ่งมีค่ามากกว่าในโมเดลตอนแรกที่ยังไม่มีการตัด Feature ที่ไม่เกี่ยวข้องออก และจะเห็นได้ว่ามีความแม่นยำมากขึ้น จากเดิม 0.63 เป็น 0.65 แต่อย่างไรก็ตาม โมเดลนี้ยังคงตอบคำถามเกี่ยวกับการทำนายลักษณะของลูกค้าที่ยังใช้บริการกับบริษัท telco ได้ดีกว่าการทำนายลักษณะของลูกค้าที่ยกเลิกการใช้บริการอยู่ดีเนื่องจากค่า precision, recall และ f1-score ที่ค่า y เป็น 0 (not churn) มีค่าที่มากกว่าค่อนข้างมาก

ตารางที่ 5 Evaluation of model

	precision	recall	f1-score
0 (not churn)	0.84	0.90	0.87
1 (churn)	0.65	0.53	0.58
accuracy		0.8031	

เมื่อพิจารณาค่า VIF อีกครั้ง ดังตารางที่ 6 จะพบว่า โมเดลนี้มีความน่าเชื่อถือและมีความแม่นยำมากขึ้น เนื่องจากค่า VIF อยู่ระหว่างค่า 1 – 3 ซึ่งนับว่าเป็นค่าความแปรปรวนที่น้อยมาก

ตารางที่ 6 Variance inflation factor

	Feature	VIF
1.	SeniorCitizen	1.2930
2.	tenure	3.0275
3.	PaperlessBilling	2.2580
4.	MultipleLines_Yes	2.2756
5.	InternetService_Fiber optic	2.2635
6.	OnlineSecurity_Yes	1.6725
7.	TechSupport_Yes	1.6988
8.	Contract_One year	1.3022

ในการทำนายโอกาสที่จะเกิดเหตุการณ์ที่สนใจ สามารถหาได้จากสมการโลจิสติค โดยที่สมการนั้นจะต้องเลือกตัวแปรที่มีความเหมาะสม เพื่อให้ความถูกต้องในการทำนายมีค่าสูงสุด จึงทำให้ได้สมการในการหาความน่าจะเป็นของการเกิดเหตุการณ์ที่สนใจ (churn) เป็นดังนี้

$$p(y) = \frac{1}{1+e^{-z}} = \frac{e^{z}}{1+e^{z}}$$

โดยที่ Z = - 0.9261 + 0.3992SeniorCitizen - 0.0433tenure + 0.5870PaperlessBilling + 0.2648MultipleLines_Yes + 1.3799InternetService_Fiber optic - 0.1986OnlineSecurity_Yes - 0.2723TechSupport_Yes - 0.4217Contract_One year

เมื่อพิจารณาตารางที่ 7 จะพบว่า การเป็นผู้สูงอายุมีอายุ 65 ปีขึ้นไป การที่ลูกค้าเลือกการเรียกเก็บค่าใช้บริการแบบลดการใช้กระดาษ การที่ลูกค้าสมัครโทรศัพท์หลายสายกับบริษัท และการลูกค้าสมัครใช้บริการอินเทอร์เน็ตผ่านโครงข่ายประเภท Fiber Optic จะเพิ่มโอกาสให้ลูกค้ายกเลิกบริการกับทางบริษัท ในขณะเดียวกัน ยิ่งระยะเวลาในการอยู่กับบริษัทมากขึ้น มีการใช้บริการการรักษาความปลอดภัยออนไลน์ การสนับสนุนด้านเทคนิค และประเภทสัญญาประเภท 1 ปี จะช่วยลดโอกาสในการที่ลูกค้าจะยกเลิกการใช้บริการกับทางบริษัทได้

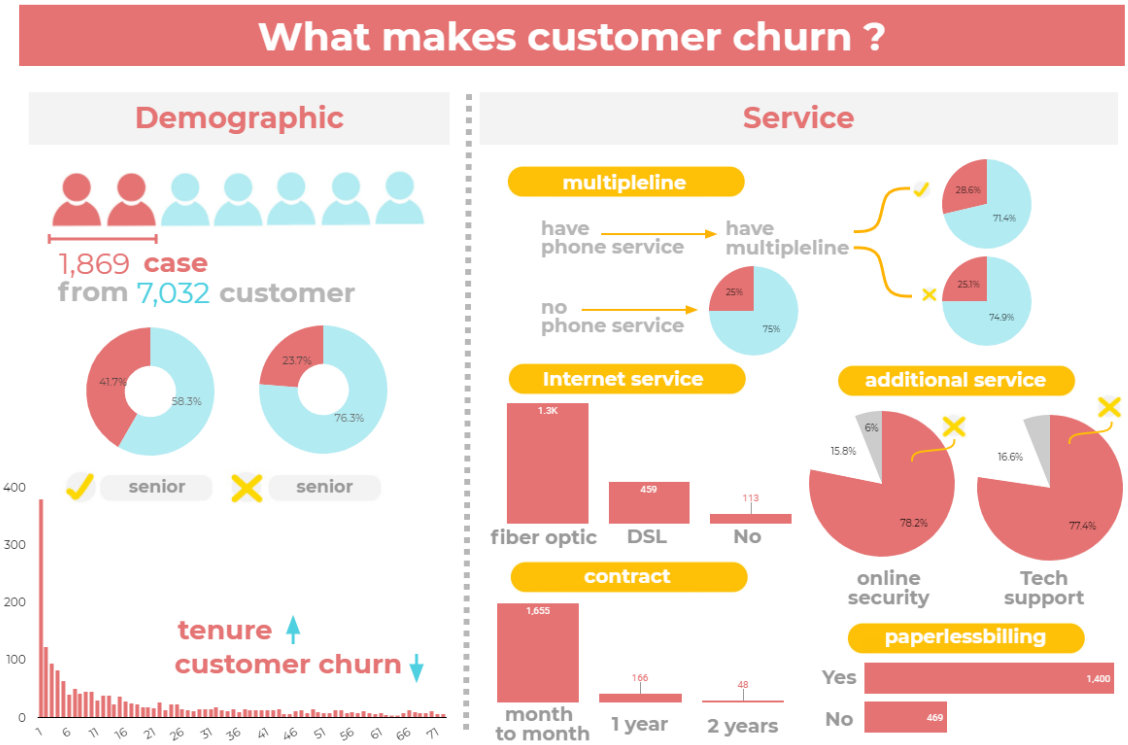
ตารางที่ 7 intercept & coefficient

	b
Constant	- 0.9261
Variable	
SeniorCitizen	0.3992
tenure	- 0.0433
PaperlessBilling	0.5870
MultipleLines_Yes	0.2648
InternetService_Fiber optic	1.3799
OnlineSecurity_Yes	- 0.1986
TechSupport_Yes	- 0.2723
Contract_One year	- 0.4217

ดูรายละเอียดเพิ่มเติมได้ที่ **model&evaluation.ipynb**

5. Deployment

แม้ว่าผลลัพธ์ที่ได้จากการทำโมเดลจะเป็นประโยชน์ แต่อย่างไรก็ตามต้องสามารถนำองค์ความรู้ที่ได้เหล่านี้ไปใช้ได้จริงได้ ดังนั้น จึงได้จัดสร้างรายงานเพื่อให้ผู้บริหารหรือนักการตลาดเข้าใจได้ง่ายและสามารถนำไปออกโปรโมชั่น หรือปรับกลยุทธ์ต่าง ๆ ที่จะช่วยให้บริษัทสามารถรักษาลูกค้าไว้ได้ ซึ่งสามารถดูรายงานได้จาก https://datastudio.google.com/reporting/c1fbceae-0b9e-4dc8-9935-f83f2dcaf9e8



รูปภาพที่ 11 ตัวอย่างรายงานแสดงผล

ด้านข้อมูลทั่วไปของลูกค้า

1. มีการสรุปลักษณะของลูกค้าในรูปแบบจำนวนและใช้ icon ประกอบ โดยที่ icon สีแดง แสดงถึงลูกค้าที่ยกเลิกใช้บริการ และสีฟ้า แสดงถึงลูกค้าที่ยังคงใช้บริการ ซึ่งการใช้ icon คนละสีจะทำให้เห็นการเปรียบเทียบที่ค่อนข้างง่าย จะเห็นได้ว่าลูกค้าที่ยกเลิกบริการมีประมาณเกือบ 1/3 ของลูกค้าทั้งหมด
2. ใช้ donut chart ในการเปรียบเทียบว่า ในกรณีที่ลูกค้าเป็นผู้สูงอายุที่มีอายุมากกว่า 65 ปี กับลูกค้าที่ไม่ได้เป็นผู้สูงอายุที่มีอายุมากกว่า 65 ปี มีอัตราการที่ลูกค้าใช้บริการต่อหรือยกเลิกบริการแตกต่างกันมากน้อยเพียงใด จะพบว่า ผู้ที่เป็นผู้สูงอายุที่มีอายุมากกว่า 65 ปี มีอัตราการ churn มากกว่า ซึ่งมากกว่าเกือบสองเท่าเลยทีเดียว
3. ใช้ bar chart เพื่อแสดงแนวโน้มของจำนวนลูกค้าที่ยกเลิกการใช้บริการกับบริษัทในระยะเวลา 72 เดือน โดยจะเห็นว่า ยิ่งลูกค้าอยู่กับบริษัทนาน จำนวนลูกค้าที่จะยกเลิกการใช้บริการก็จะมีน้อยลง จาก bar chart จะเห็นชัดเลยว่าระยะเวลา 1 เดือนจำนวนลูกค้าที่ยกเลิกบริการมีสูงมาก ดังนั้นทางบริษัทควรมีข้อเสนอ หรือดูแลหลังการขายให้กับลูกค้าเป็นอย่างดี เพื่อให้ลูกค้าไม่เปลี่ยนใจยกเลิกการใช้บริการ

ด้านการบริการต่าง ๆ

1. ในด้าน phone service หากลูกค้ามีบริการ phone service กับทางบริษัท จะส่งผลต่อบริการ multiple lines ซึ่งในส่วนนี้จะใช้ pie chart ในการเปรียบเทียบ จะเห็นได้ว่าการที่ลูกค้าสมัครโทรศัพท์หลายสายกับบริษัทจะทำให้อัตราที่ลูกค้าที่จะยกเลิกการใช้บริการมีมากกว่าการที่ไม่มีการสมัครใช้บริการโทรศัพท์หลายสายกับบริษัท และไม่มีบริการทางด้านโทรศัพท์
2. บริการด้าน Internet ในส่วนนี้จะแสดงเฉพาะลูกค้าที่ยกเลิกการใช้บริการเท่านั้น เนื่องจากจำนวนลูกค้าที่ไม่ได้ยกเลิกการใช้บริการในด้าน internet มีจำนวนที่ใกล้เคียงกันและในการวิเคราะห์ครั้งนี้จะเน้นหา insight ของกลุ่มลูกค้าที่ยกเลิกการใช้บริการ ดังนั้นจึงขอไม่แสดงในส่วนนี้ โดยในส่วนนี้จะใช้ bar chart ในการแสดงผล จะเห็นได้ว่าโครงข่าย internet ประเภท fiber optic ทำให้อัตราการยกเลิกใช้บริการของลูกค้ามีสูงมาก ดังนั้นบริษัทควรที่จะทำการทบทวนเกี่ยวกับโครงข่าย internet ประเภท fiber optic เพื่อหาข้อบกพร่องที่เกิดขึ้น ซึ่งถ้าหากแก้ไขได้ อาจจะส่งผลให้จำนวนลูกค้าที่ยกเลิกการใช้บริการลดลง
3. การไม่มีการคุ้มครอง ไม่ว่าจะเป็นบริการด้านการคุ้มครองความปลอดภัยออนไลน์หรือว่าการมีทีมสนับสนุนทางด้านเทคนิค จะทำให้อัตราการยกเลิกใช้บริการของลูกค้ามีสูงมาก ดังนั้นทางบริษัทอาจจะต้องมีการนำเสนอในส่วนของด้านการคุ้มครองให้กับลูกค้าอย่างชัดเจน เพื่อให้ลูกค้าได้ทราบว่ามีการคุ้มครองอะไรบ้าง ซึ่งอาจจะส่งผลให้อัตราการยกเลิกบริการของลูกค้าลดลงได้ ในส่วนนี้จะแสดงเฉพาะลูกค้าที่ยกเลิกการใช้บริการเท่านั้นเช่นกัน และจะแสดงผลโดยการใช้ pie chart
4. ด้านสัญญาในการใช้บริการ จะแสดงเฉพาะลูกค้าที่ยกเลิกการใช้บริการและใช้ bar chart ในการแสดงผล จะเห็นได้ชัดเลยว่าสัญญาประเภทเดือนต่อเดือน ทำให้อัตราการที่ลูกค้ายกเลิกการใช้บริการมีสูงมากสูงถึง 1,655 คน จาก 1,869 ซึ่งสอดคล้องกับค่าสัมประสิทธิ์ของ Contract_One year ที่มีค่าเป็นลบ ที่ทำให้เห็นว่าการทำสัญญาประเภท 1 ปี จะช่วยลดโอกาสที่ลูกค้าจะยกเลิกการใช้บริการได้ ดังนั้นทางบริษัทอาจจะมีการโน้มน้าวหรือมีข้อเสนอพิเศษต่าง ๆ เพื่อให้ลูกค้าต่อสัญญากับเราในระยะเวลาที่นานมากขึ้น
5. การที่ลูกค้ามีการเลือกใช้การเรียกเก็บค่าใช้บริการแบบลดการใช้กระดาษทำให้อัตราการยกเลิกการใช้บริการมีสูงกว่าการไม่ลดการใช้กระดาษ ดังนั้นทางบริษัทจึงควรหาสาเหตุว่าหากลูกค้าที่เลือกการเก็บค่าบริการแบบลดการใช้กระดาษ มีวิธีการชำระเงินแบบไหนที่ทำให้เกิดข้อบกพร่องมากที่สุด และทำการแก้ไขข้อบกพร่องจุดนี้ โดยในส่วนนี้จะแสดงผลโดยใช้ bar chart ในแนวนอน และจะแสดงเฉพาะลูกค้าที่ยกเลิกการใช้บริการเช่นเดียวกัน

ข้อมูลที่แสดงบน Google Data Studio ได้มาจากการใช้ bigquery ในการเลือกข้อมูลเพื่อนำมาแสดงผล จะนำข้อมูล data_begin จาก cloud storage มาสร้าง table ชื่อว่า data_begin ซึ่งข้อมูลในส่วนนี้เป็นข้อมูลที่ทำการ drop missing value และ เปลี่ยน data type ให้ถูกต้องเท่านั้น ซึ่งอยู่ในขั้นตอน data preparation

ซึ่งจะประกอบไปด้วย 9 ตาราง ดังนี้

- Churn

```
SELECT
    Churn,
    count(Churn) AS num
FROM
    churn.data_begin
GROUP BY Churn
```

- SeniorCitizen

```
SELECT
    Churn,
    SUM(CASE WHEN SeniorCitizen = 0 THEN 1 END) AS noSenior,
    SUM(CASE WHEN SeniorCitizen = 1 THEN 1 END) AS isSenior
FROM
    `churn.data_begin`
GROUP BY Churn
```

- tenure

```
SELECT
    tenure,
    count(tenure) AS num
FROM
    churn.data_begin
WHERE Churn=true
GROUP BY tenure
ORDER BY tenure
```

- MultipleLines

```
SELECT
    Churn,
    SUM(CASE WHEN MultipleLines = 'Yes' THEN 1 END) AS isYes,
    SUM(CASE WHEN MultipleLines = 'No' THEN 1 END) AS isNo,
    SUM(CASE WHEN MultipleLines = 'No phone service' THEN 1 END) AS isNophone,
FROM
    churn.data_begin
GROUP BY Churn
```

- InternetService

```
SELECT
    InternetService,
    count(InternetService)
FROM
    churn.data_begin
WHERE Churn=true
GROUP BY InternetService
```

- OnlineSecurity

```
SELECT
    OnlineSecurity,
    count(OnlineSecurity)
FROM
    churn.data_begin
WHERE Churn=true
GROUP BY OnlineSecurity
```

- TechSupport

```
SELECT
    TechSupport,
    count(TechSupport)
FROM
    churn.data_begin
WHERE Churn=true
GROUP BY TechSupport
```

- Contract

```
SELECT
    Contract,
    count(Contract)
FROM
    churn.data_begin
WHERE Churn=true
GROUP BY Contract
```

- Paperlessbilling

```
SELECT
    PaperlessBilling,
    count(PaperlessBilling)
FROM
    churn.data_begin
WHERE Churn=true
GROUP BY PaperlessBilling
```

- **สรุปผลการศึกษา อภิปรายผล และข้อเสนอแนะ**

การที่ลูกค้าเป็นผู้สูงอายุมีอายุ 65 ปีขึ้นไป จะเพิ่มโอกาสให้ลูกค้ายกเลิกการใช้บริการกับทางบริษัท อาจจะด้วยเหตุผลที่ว่าผู้สูงอายุในช่วง 65 ปีขึ้นไป อยู่ในวัยเกษียณ อาจจะไม่มีความจำเป็นที่จะต้องใช้บริการ หรืออาจจะขาดการประชาสัมพันธ์ด้านการเข้าถึงอินเทอร์เน็ต รวมถึงอุปสรรคในการการเข้าถึงอินเทอร์เน็ต ได้แก่ ค่าบริการอินเทอร์เน็ตมีราคาสูง

การที่ลูกค้าสมัครใช้บริการอินเทอร์เน็ตผ่านโครงข่ายประเภท Fiber Optic จะเพิ่มโอกาสให้ลูกค้ายกเลิกการใช้บริการกับทางบริษัท นับว่าเป็นปัจจัยที่สำคัญต้น ๆ ที่ทำให้อัตราการยกเลิกการใช้บริการสูงมาก

การที่ลูกค้าเลือกการเรียกเก็บค่าใช้บริการแบบลดการใช้กระดาษ และการที่ลูกค้าสมัครโทรศัพท์หลายสายกับบริษัท มีผลต่อการยกเลิกการใช้บริการของลูกค้า

ระยะเวลาในการอยู่กับบริษัทนับว่าเป็นปัจจัยสำคัญที่สำคัญที่สุดในการยกเลิกการใช้บริการ ยิ่งลูกค้าอยู่กับบริษัทได้นานมากเท่าไหร่ก็จะลดโอกาสในการยกเลิกการใช้บริการมากขึ้นเท่านั้น และจะเห็นได้ชัดว่าส่วนใหญ่แล้วลูกค้าจะยกเลิกการใช้บริการภายใน 1 เดือนในอัตราที่สูงมาก ซึ่งอาจจะสอดคล้องกับสัญญาด้านการให้บริการ หากเป็นสัญญาประเภท 1 ปี จะช่วยลดโอกาสในการที่ลูกค้าจะยกเลิกการใช้บริการกับทางบริษัท แต่หากเป็นสัญญาประเภทเดือนต่อเดือน จะส่งผลให้เพิ่มโอกาสในการที่ลูกค้าจะยกเลิกการใช้บริการกับทางบริษัท

ทางบริษัทควรแนะนำบริการเสริมทางด้านโทรศัพท์และอินเตอร์เน็ต เช่น บริการการรักษาความปลอดภัยออนไลน์และการสนับสนุนด้านเทคนิค เนื่องจากจะช่วยลดโอกาสในการที่ลูกค้าจะยกเลิกการใช้บริการกับทางบริษัทได้

การใช้เทคนิคการวิเคราะห์แบบ Logistic Regression สามารถคำนวณหาความน่าจะเป็นโอกาสในการที่ลูกค้าจะยกเลิกการใช้บริการ ได้ดังนี้

$$p(y) = \frac{1}{1+e^{-z}} = \frac{e^{z}}{1+e^{z}}$$

โดยที่ Z = - 0.9261 + 0.3992SeniorCitizen - 0.0433tenure + 0.5870PaperlessBilling + 0.2648MultipleLines_Yes + 1.3799InternetService_Fiber optic - 0.1986OnlineSecurity_Yes - 0.2723TechSupport_Yes - 0.4217Contract_One year

ซึ่งโมเดลนี้ให้ค่า accuracy score เท่ากับ 0.8031 และมีความน่าเชื่อถือและมีความแม่นยำมากขึ้น เนื่องจากค่า VIF อยู่ระหว่างค่า 1 – 3 ซึ่งนับว่าเป็นค่าความแปรปรวนที่น้อยมาก

ประโยชน์ในการวิเคราะห์ครั้งนี้ จะส่งผลดีกับอุตสาหกรรมด้านโทรคมนาคม เนื่องจากจะทำให้เห็นถึง insight ของลูกค้า ทำให้สามารถปรับกลยุทธ์ ที่จะช่วยลดการยกเลิกการใช้บริการของลูกค้าได้ รวมถึงช่วยลดต้นทุนในการหาลูกค้าใหม่ได้ เนื่องจากสามารถรู้เท่าทันถึงพฤติกรรมของลูกค้าที่จะยกเลิกการใช้บริการ

นอกจากนี้ในการศึกษาครั้งนี้ทำให้เห็นว่าโมเดลนี้ตอบคำถามเกี่ยวกับการทำนายลักษณะของลูกค้าที่ยังใช้บริการกับบริษัท telco ได้ดีกว่าการทำนายลักษณะของลูกค้าที่ยกเลิกการใช้บริการ สามารถดูได้จากค่า precision, recall และ f1-score ซึ่งอาจจะเป็นผลมาจาก target variable ที่มี 2 class ไม่สมดุล มีความเอนเอียง หรืออาจจะมีเทคนิคในการวิเคราะห์ที่เหมาะสมมากกว่า Logistic Regression ดังนั้น ในการวิเคราะห์ครั้งต่อไปควรใช้หลาย ๆ เทคนิคในการวิเคราะห์ และนำผลลัพธ์ที่ได้มาเปรียบเทียบกัน เพื่อหาวิธีที่เหมาะสมที่สุดสำหรับการวิเคราะห์ข้อมูลชุดนี้ รวมไปถึงอาจจะใช้ข้อมูลที่เยอะขึ้นมากกว่านี้ เพื่อให้โมเดลสามารถตอบปัญหาเกี่ยวกับการทำนายลักษณะของลูกค้าในบริษัท Telco ได้ทั้งลูกค้าที่ยังใช้บริการและยกเลิกการใช้บริการกับบริษัท

อย่างไรก็ตามในการศึกษาครั้งนี้ดำเนินการผ่าน Google Cloud Platform ทำให้เรียนรู้ว่าเป็นแพลตฟอร์มที่มีประโยชน์มาก รองรับข้อมูลขนาดใหญ่ มีความสามารถในการวิเคราะห์และจัดการข้อมูล ทำให้ตอบโจทย์การทำงานเชิงธุรกิจได้เป็นอย่างดี รวมถึงมีบริการที่แยกย่อยออกไปอีกมากมายให้ได้เลือกใช้งาน ซึ่งในอนาคตหากเป็นไปได้อยากจะลองใช้บริการอื่น ๆ บน Google Cloud Platform อย่างเช่น Dataflow, Dataproc หรือ Dataprep ที่มี environment เหมาะแก่การจัดการกับข้อมูลให้มากกว่านี้