DATA SCIENCE ACADEMY

CAPSTONE PROJECT

RIDE HAILING INTERNET PACKAGE

GROUP 12

Yustinus Kunta Wibisana
Shelby Marsa Istiqomah
Wahyu Sejati Roso
Rizaldy Al Kautsar Utomo

CRISP-DM

RIDE-HAILING INTERNET PACKAGE

1 **Business Understanding**
Understand objective and requirement from business perspective

2 **Data Understanding**
Getting familiar with the data to form hypotheses

3 **Data Preparation**
Construct dataset from raw data

4 **Modelling**
Building the model for desired output

5 **Evaluation**
Assess the quality of the model based on requirement

6 **Deployment**
Put the result to work and achieve the goals

BUSINESS UNDERSTANDING

Ride-Hailing is **multi billion business** in Indonesia, and the driver stand as partner which plays crucial role, represent **0.9% of Telkomsel Population**

Active Ride-Hailing Drivers

1.48 Million

Serving

Driver to Consumer Ratio

1:11

Active Ride-Hailing Users

16.86 Million

Gojek drivers as a partner contribute to **IDR 8 Trillion annually (2018)**

34% of them have monthly income of > IDR 3.5 million after joined as ride hailing driver, only 8% of them already have it before join as a partner

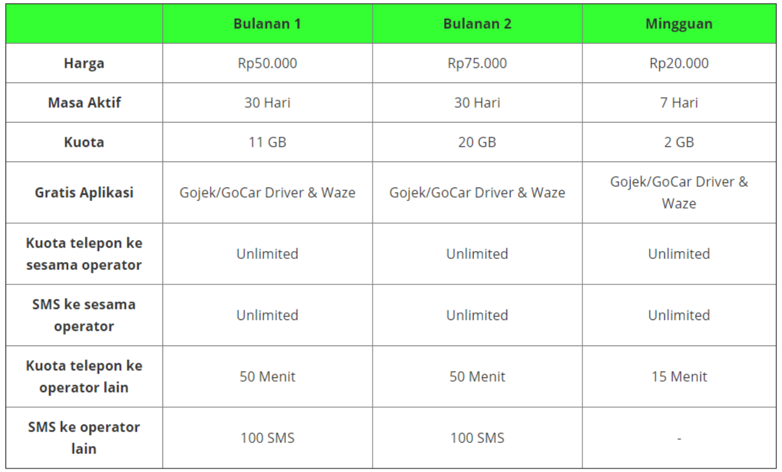
Daily income of IDR 150-200k/day, communication expense can be a burden if **Telco** company don't provide the **best offering**

*Telkomsel MSIGHT January 2020 Data

** Lembaga Demografi FEB UI (2018)

Telkomsel as connectivity provider stand as **enabler** to make the **driver experience with ride-hailing app** more **seamless**

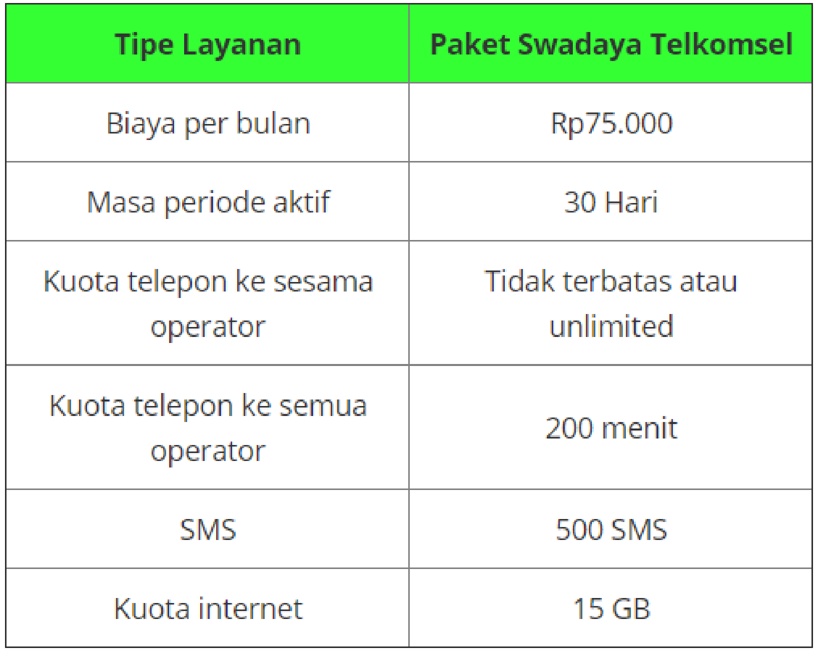
XL



	Bulanan 1	Bulanan 2	Mingguan
Harga	Rp50.000	Rp75.000	Rp20.000
Masa Aktif	30 Hari	30 Hari	7 Hari
Kuota	11 GB	20 GB	2 GB
Gratis Aplikasi	Gojek/GoCar Driver & Waze	Gojek/GoCar Driver & Waze	Gojek/GoCar Driver & Waze
Kuota telepon ke sesama operator	Unlimited	Unlimited	Unlimited
SMS ke sesama operator	Unlimited	Unlimited	Unlimited
Kuota telepon ke operator lain	50 Menit	50 Menit	15 Menit
SMS ke operator lain	100 SMS	100 SMS	-

Price/GB	IDR 3750

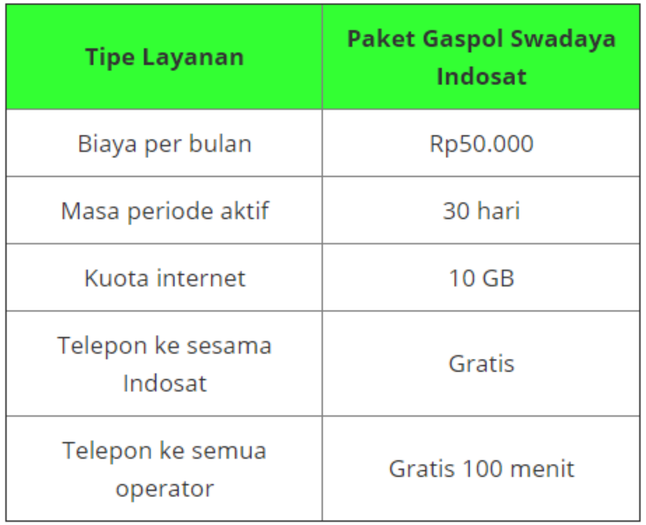
Telkomsel



Tipe Layanan	Paket Swadaya Telkomsel
Biaya per bulan	Rp75.000
Masa periode aktif	30 Hari
Kuota telepon ke sesama operator	Tidak terbatas atau unlimited
Kuota telepon ke semua operator	200 menit
SMS	500 SMS
Kuota internet	15 GB

Price/GB	IDR 5000

Indosat

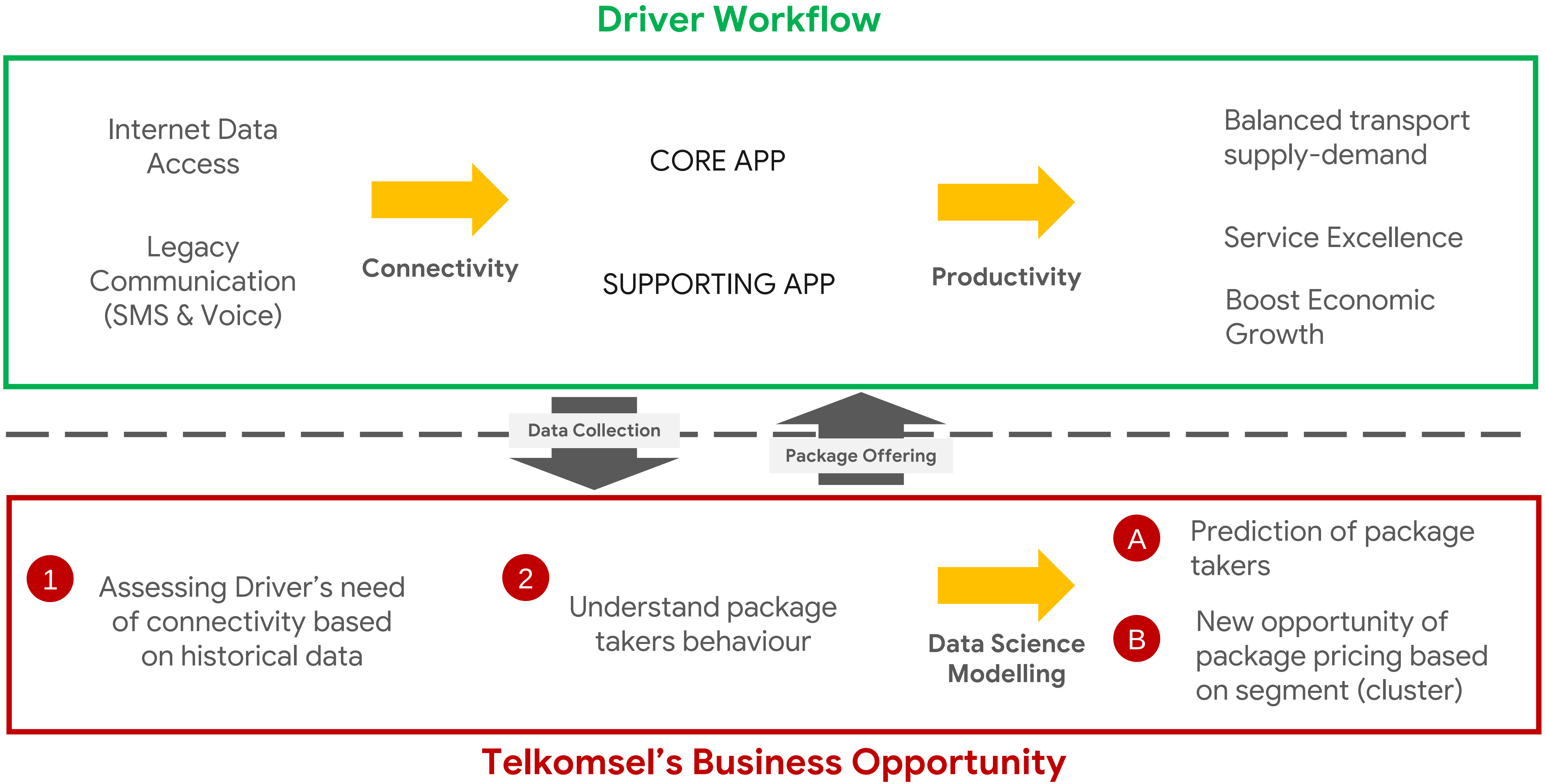


Tipe Layanan	Paket Gaspol Swadaya Indosat
Biaya per bulan	Rp50.000
Masa periode aktif	30 hari
Kuota internet	10 GB
Telepon ke sesama Indosat	Gratis
Telepon ke semua operator	Gratis 100 menit

Price/GB	IDR 5000

Telkomsel has several competitor with **more competitive price for ride-hailing driver package.** We need to **enhance value proposition** to improve the takers of package

Understanding **driver needs and usage of connectivity** with **data science** can be the key to drive **more takers in Ride-Hailing package**



Driver Workflow

Internet Data Access

Legacy Communication (SMS & Voice)

Connectivity

CORE APP

SUPPORTING APP

Productivity

Balanced transport supply-demand

Service Excellence

Boost Economic Growth

Data Collection

Package Offering

1 Assessing Driver's need of connectivity based on historical data

2 Understand package takers behaviour

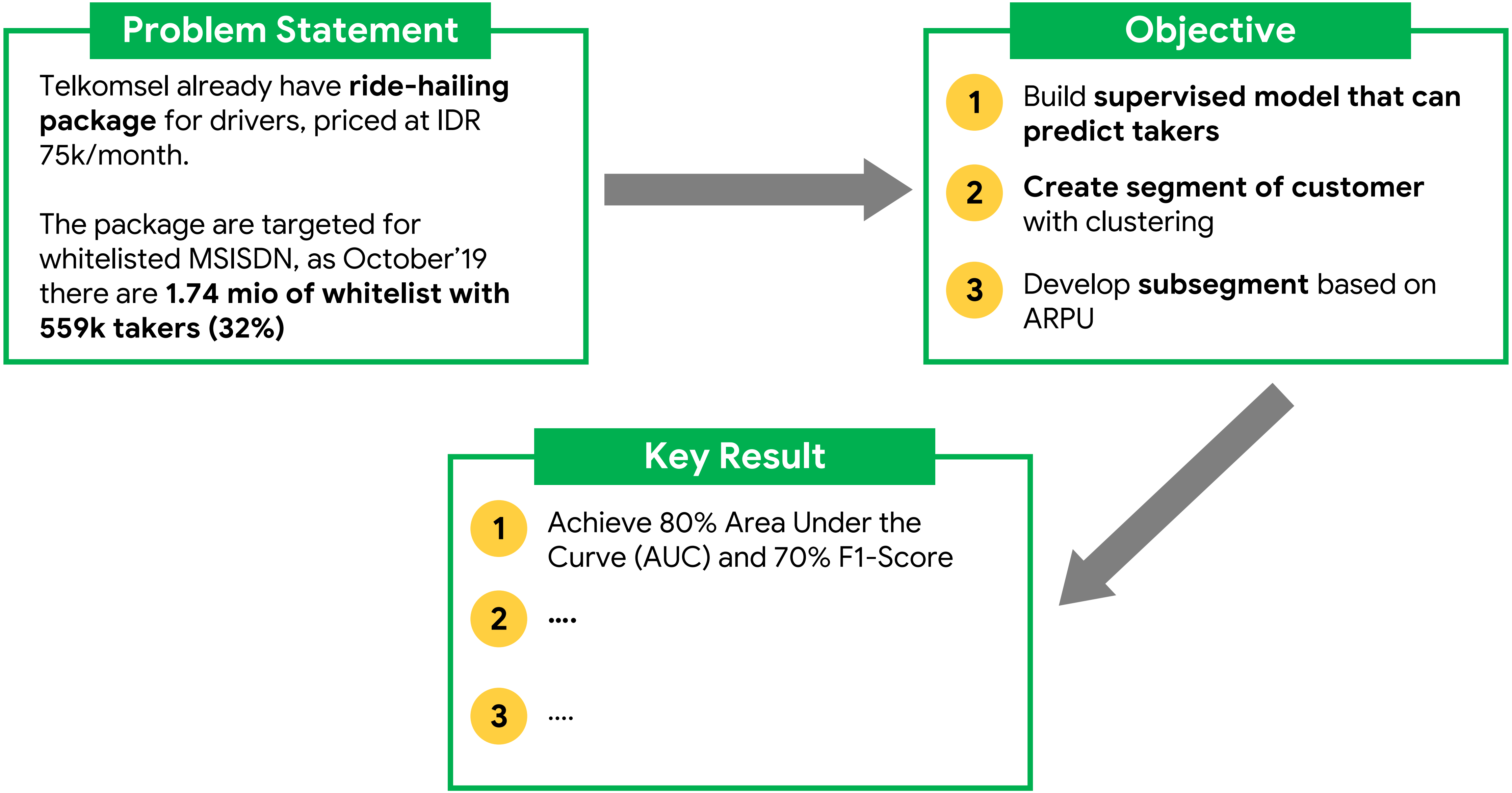
Data Science Modelling

A Prediction of package takers

B New opportunity of package pricing based on segment (cluster)

Telkomsel's Business Opportunity

... based on previous business problem, **3** objective and key result can be derived wuith data science process (classification & clustering)



Problem Statement

Telkomsel already have **ride-hailing package** for drivers, priced at IDR 75k/month.

The package are targeted for whitelisted MSISDN, as October'19 there are **1.74 mio of whitelist with 559k takers (32%)**

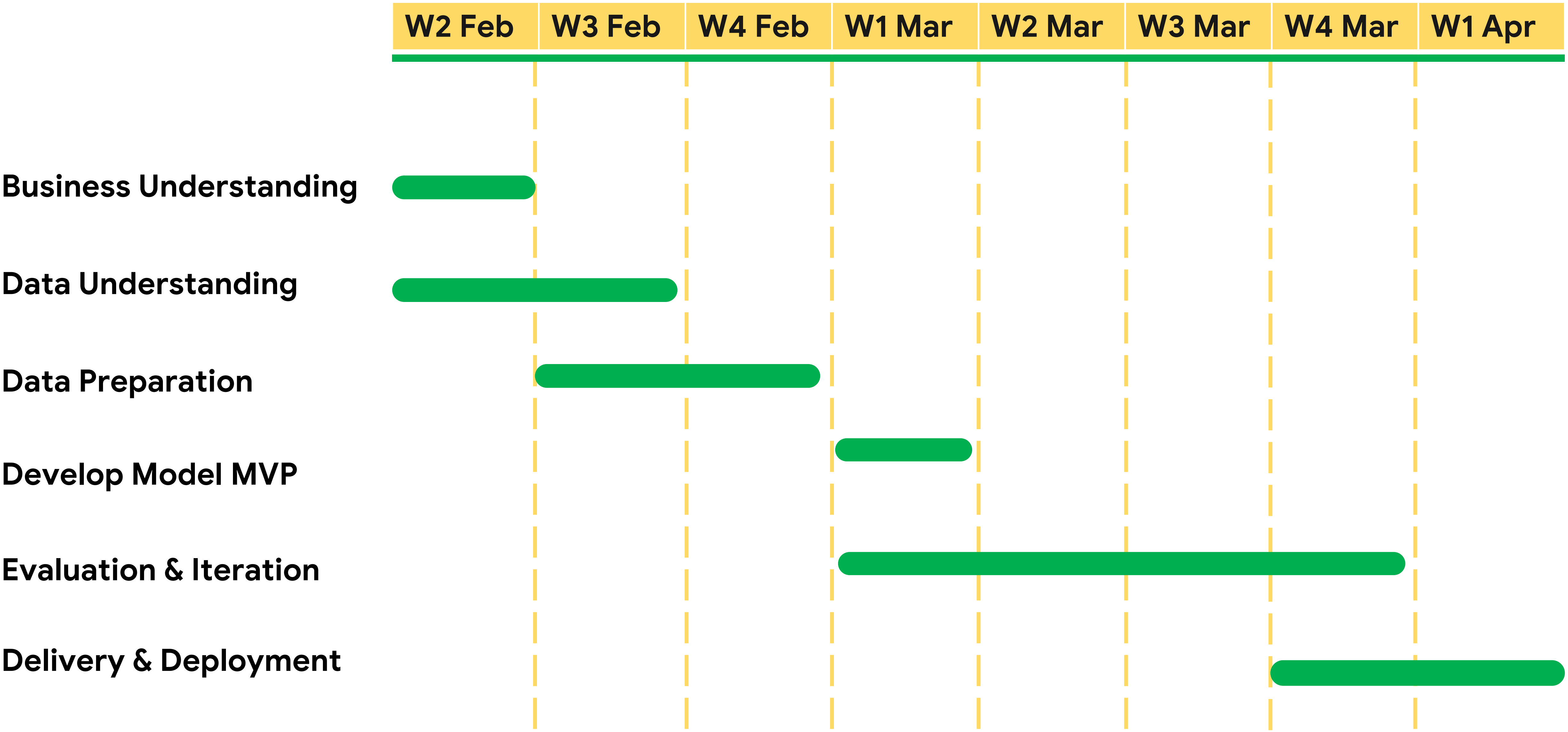
Objective

1. Build **supervised model that can predict takers**
2. **Create segment of customer** with clustering
3. Develop **subsegment** based on ARPU

Key Result

1. Achieve 80% Area Under the Curve (AUC) and 70% F1-Score
2. ....
3. ....

… to achieve the result, expected timeline is **2 months** of CRISP-DM complete cycle

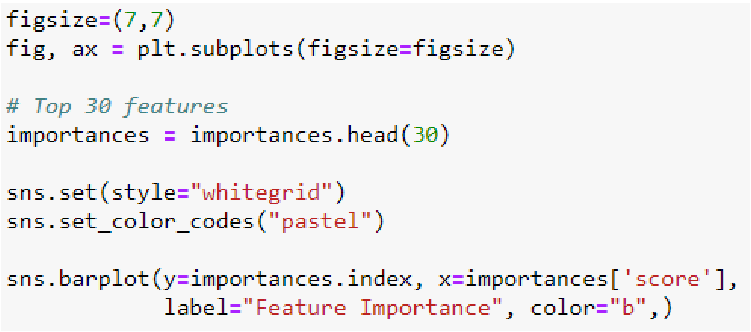


DATA UNDERSTANDING

DATA PREPARATION

MODELLING

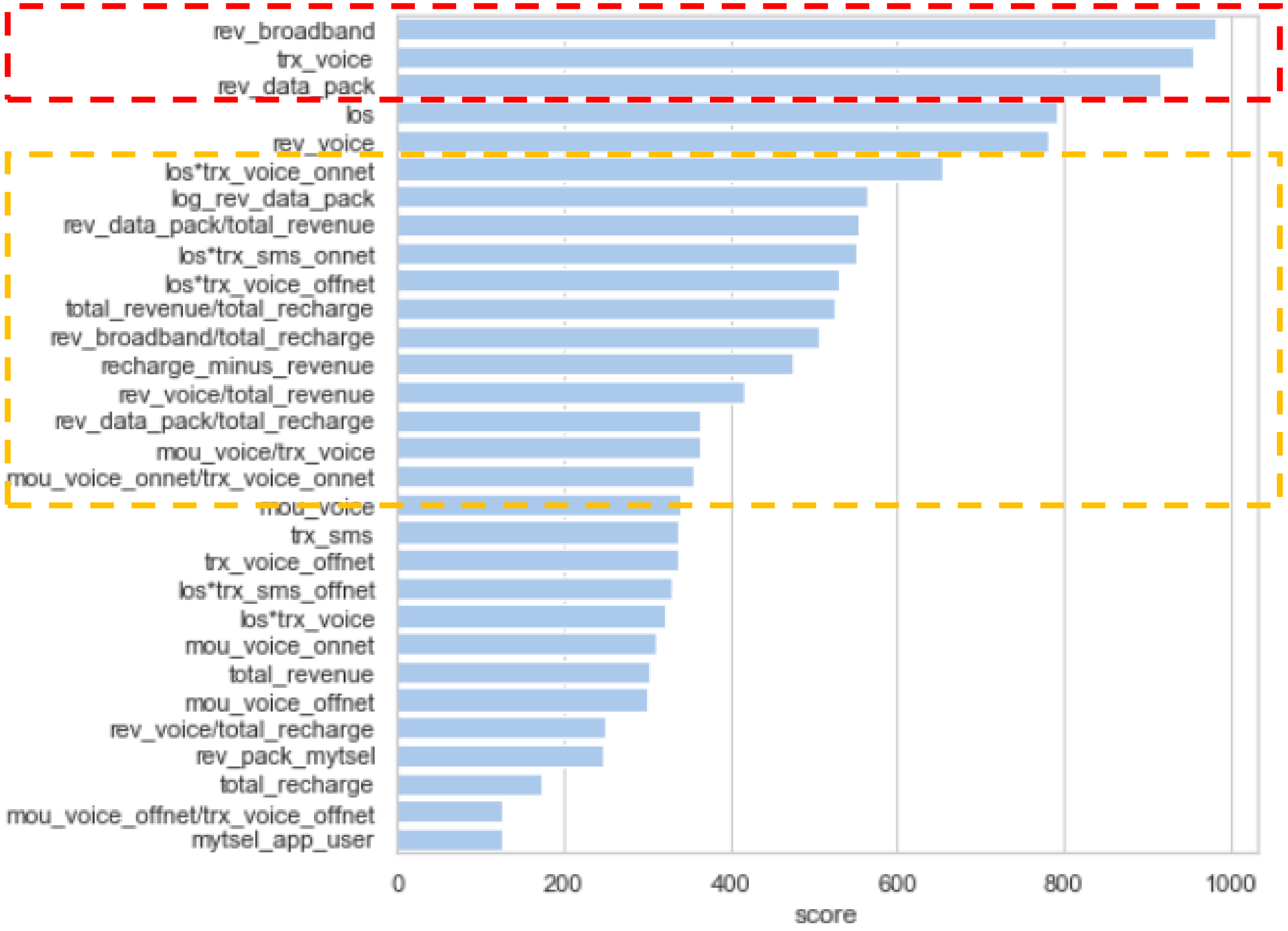
30 top feature ingested into the model**, feature selection** conducted using **XGBoost algorithm**



```
figsize=(7,7)
fig, ax = plt.subplots(figsize=figsize)

# Top 30 features
importances = importances.head(30)

sns.set(style="whitegrid")
sns.set_color_codes("pastel")

sns.barplot(y=importances.index, x=importances['score'],
            label="Feature Importance", color="b",)
```

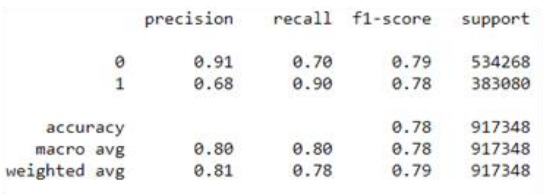
Revenue of **data usage (rev_broadband & rev_data_pack)** managed to get into **top 3**

Feature that created from **feature engineering** managed to have high score

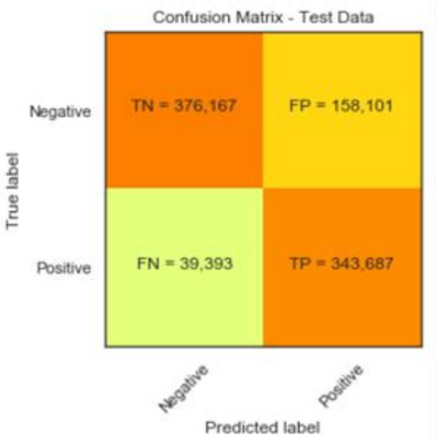


For 1st Objective (Classification), Random Forest achieved highest score, compared to Logistic Regression and Decision Tree

Logistic Regression

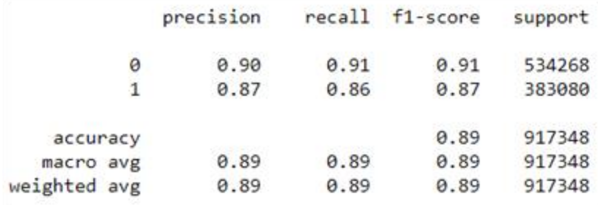


```
              precision    recall  f1-score   support

           0       0.91      0.70      0.79    534268
           1       0.68      0.90      0.78    383080

    accuracy                           0.78    917348
   macro avg       0.80      0.80      0.78    917348
weighted avg       0.81      0.78      0.79    917348
```

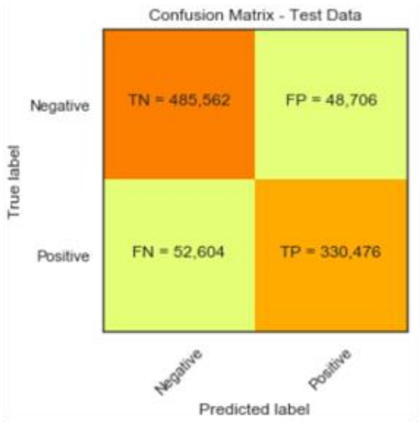


F1 Score= 79%

Decision Tree

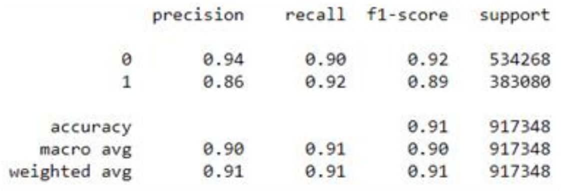


```
              precision    recall  f1-score   support

           0       0.90      0.91      0.91    534268
           1       0.87      0.86      0.87    383080

    accuracy                           0.89    917348
   macro avg       0.89      0.89      0.89    917348
weighted avg       0.89      0.89      0.89    917348
```

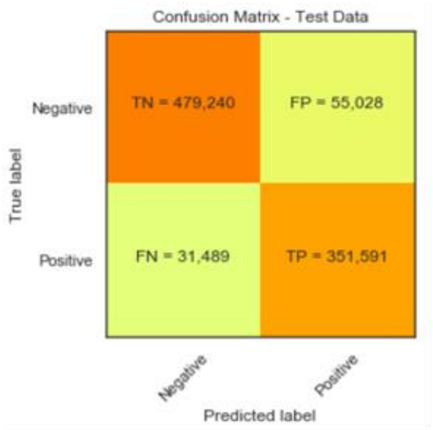


F1 Score= 89%

Random Forest



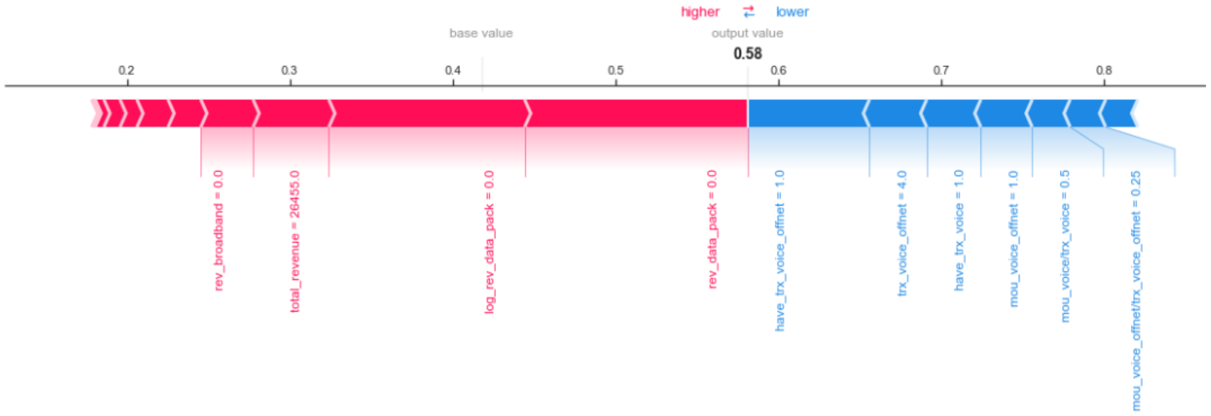
```
              precision    recall  f1-score   support

           0       0.94      0.90      0.92    534268
           1       0.86      0.92      0.89    383080

    accuracy                           0.91    917348
   macro avg       0.90      0.91      0.90    917348
weighted avg       0.91      0.91      0.91    917348
```

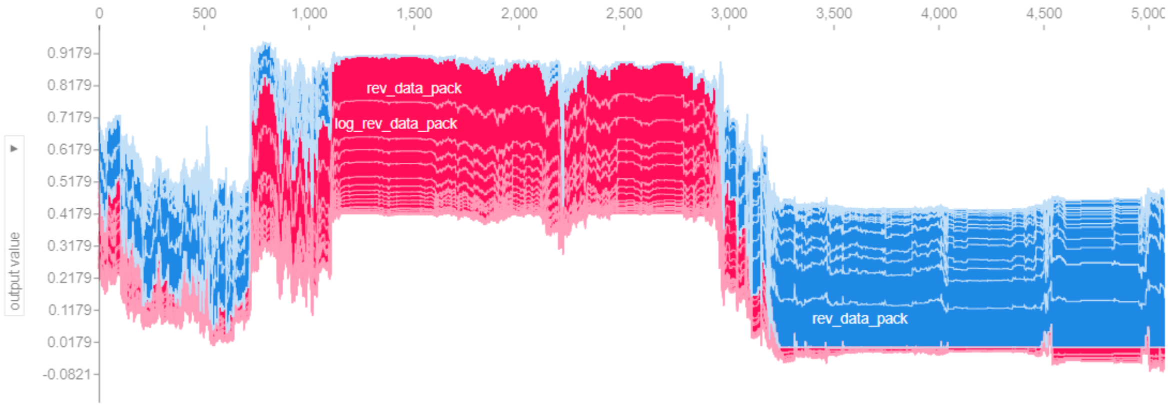


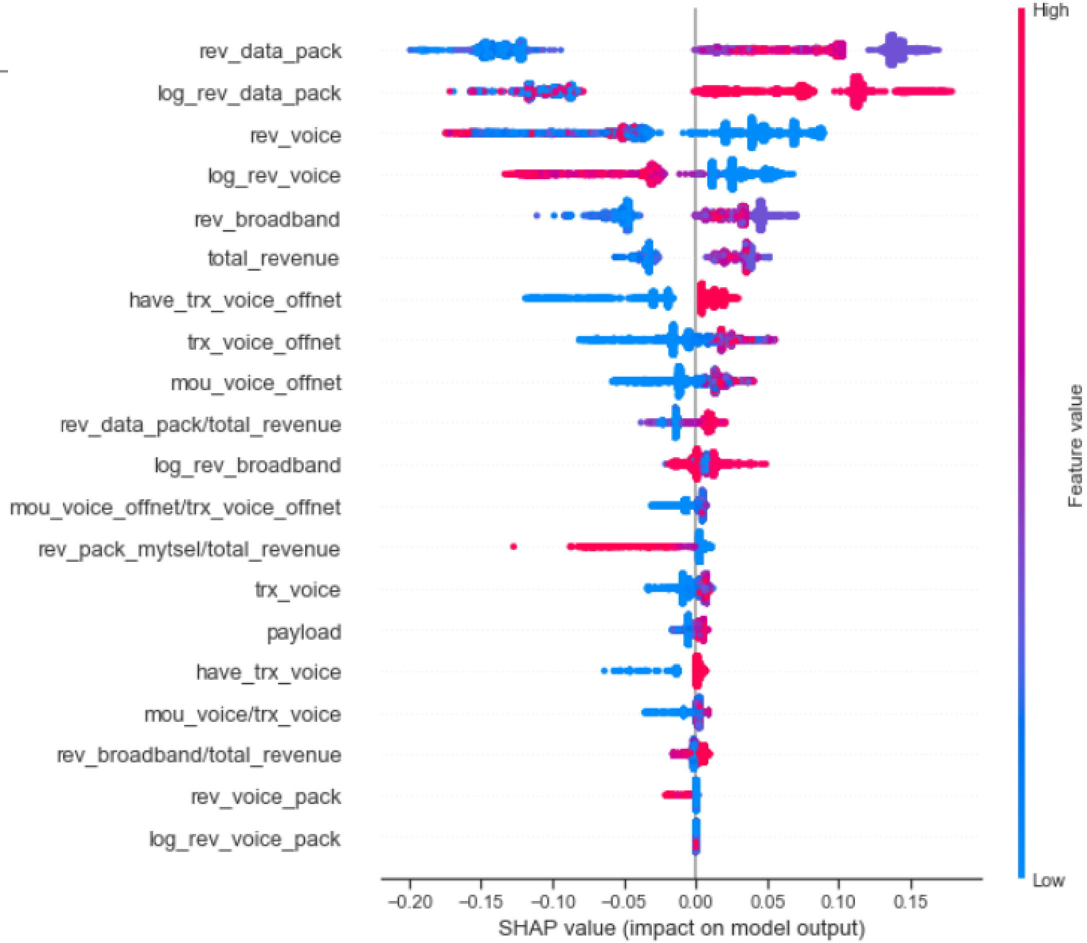
F1 Score= 91%

All algorithm managed to achieve **key result of F1-Score above 70%,** we decided to went with **Random Forest** that managed to have high precision and recall resulting with **high F1-Score (91%)**

SHAP Value also determined that **data_package** is the most important feature, followed by **voice**

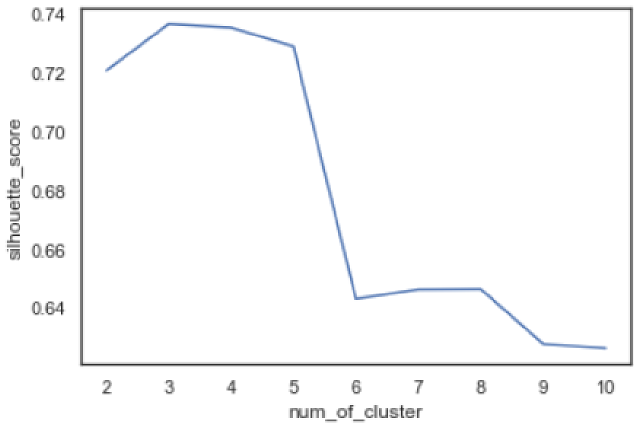


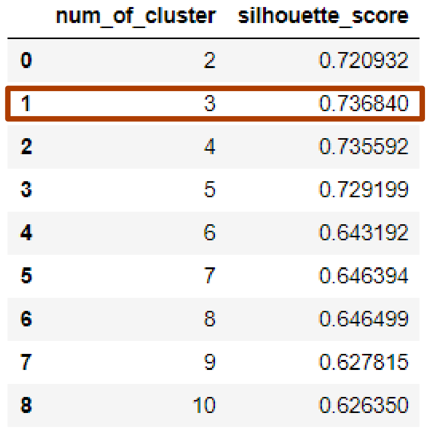




For 2nd Objective (Clustering), **silhouette score** is used to determine number of cluster. **The optimal number of cluster is 3**

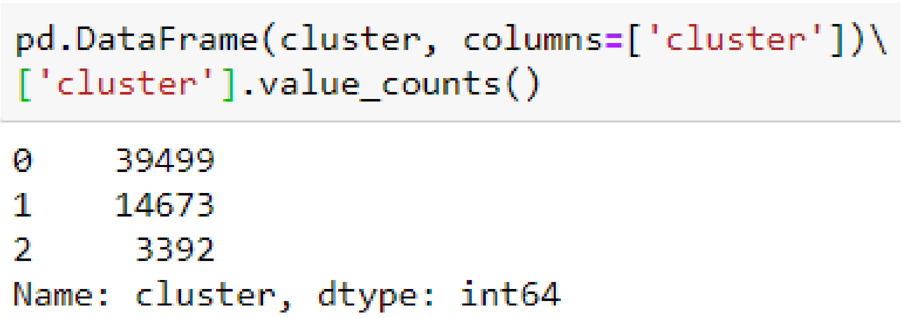
Determining number of cluser





	num_of_cluster	silhouette_score
0	2	0.720932
1	3	0.736840
2	4	0.735592
3	5	0.729199
4	6	0.643192
5	7	0.646394
6	8	0.646499
7	9	0.627815
8	10	0.626350

Total MSISDN in each cluster

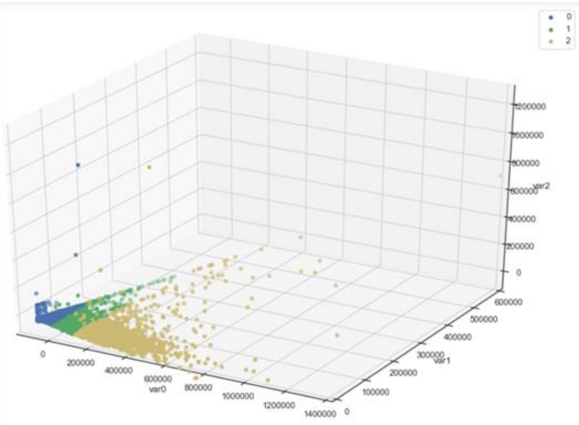


```
pd.DataFrame(cluster, columns=['cluster'])\
['cluster'].value_counts()
```

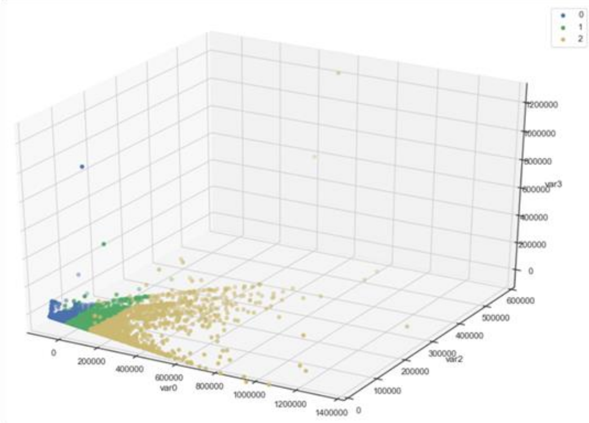
```
0    39499
1    14673
2     3392
Name: cluster, dtype: int64
```

3D Cluster Visualization using PCA Analysis

Feature 1, 2 & 3



Feature 1, 2 & 4

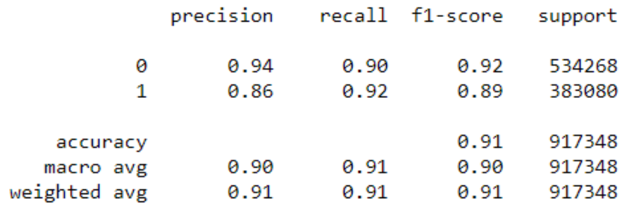


Cluster seems have consistent grouping (not underfit/overfit) based on 3D visualization.

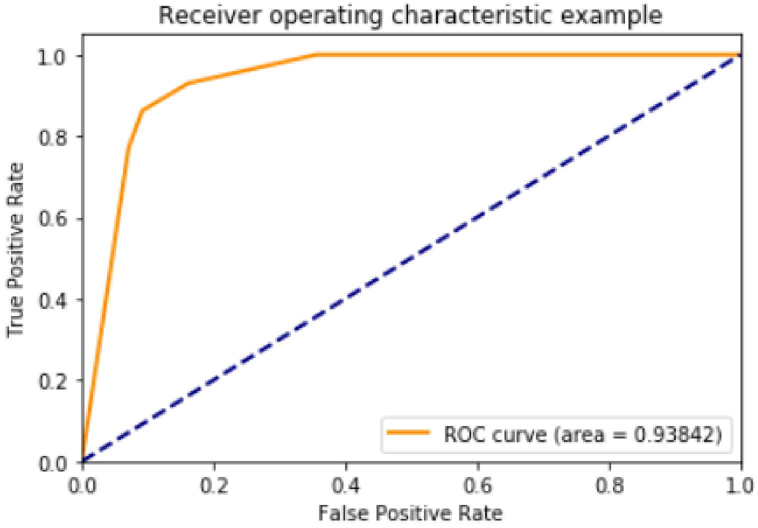
Cluster 1 have the largest number of MSISDN (39k), followed by cluster 2 (14k) and cluster 3 (3k)

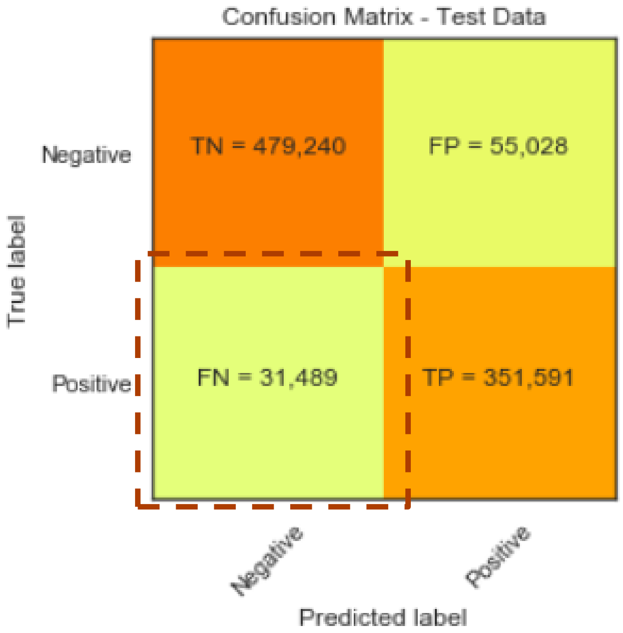
EVALUATION

Random Forest successfully meet the **1st objective for classification with 93.8% of AUC and 91% of F1-Score**



```
              precision    recall  f1-score   support

           0       0.94      0.90      0.92    534268
           1       0.86      0.92      0.89    383080

    accuracy                           0.91    917348
   macro avg       0.90      0.91      0.90    917348
weighted avg       0.91      0.91      0.91    917348
```





There are potential **31.4k new numbers of takers** based on this algorithm. In total there are **383k potential package takers.**

39.3% taker rate, uplift +7% from previous data

For the 2nd Objective, there are 3 main cluster with different behaviour and usage, resulting in different package price

Low Transaction

39.4k subs
IDR 86k
14 GB

Potential Customer

14.6k subs
IDR 164k
24 GB

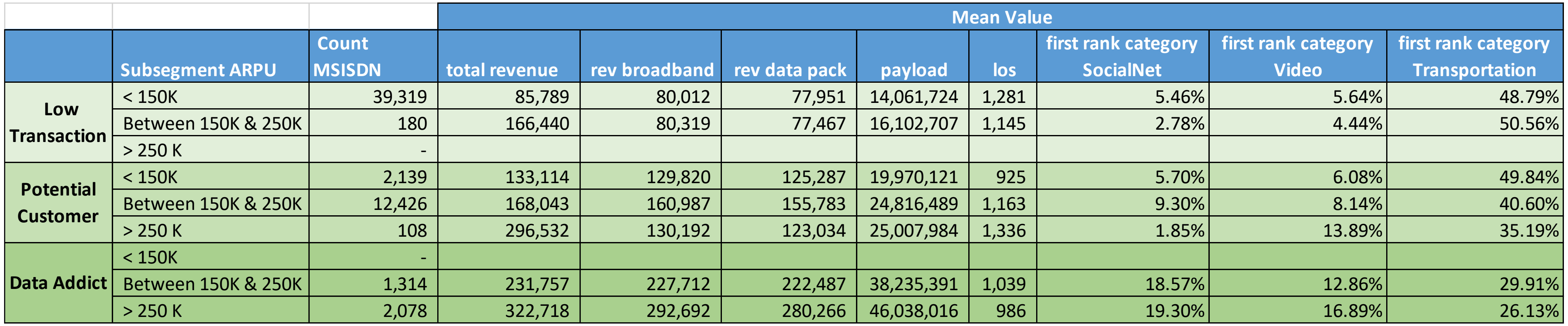
Data Addict

3.39k subs
IDR 287k
43 GB

Cluster Name	Count MSISDN	Value	total revenue	rev broadband	rev data pack	payload	los	first rank category SocialNet	first rank category Video	first rank category Transportation
Low Transaction	**39,499**	**mean**	86,156	80,014	77,949	14,071,025	1,280	5.451%	5.631%	48.801%
		25%	75,000	75,000	75,000	9,504,591	426	-	-	-
		50%	79,274	75,000	75,000	14,521,546	781	-	-	-
		75%	93,202	79,415	75,000	17,777,930	1,595	-	-	1
Potential Customer	**14,673**	**mean**	163,897	156,217	151,097	24,111,405	1,129	8.724%	7.878%	41.907%
		25%	150,000	150,000	150,000	17,945,584	381	-	-	-
		50%	155,045	150,062	150,000	23,050,458	703	-	-	-
		75%	175,230	165,118	150,010	29,545,320	1,492	-	-	1
Data Addict	**3,392**	**mean**	287,481	267,520	257,883	43,015,418	1,006	19.015%	15.330%	27.594%
		25%	234,979	225,089	225,000	33,768,140	341	-	-	-
		50%	262,002	249,963	235,000	40,680,318	646	-	-	-
		75%	307,985	293,633	280,000	49,658,070	1,274	-	-	1

For the 3rd **Objective, main cluster derived into each 3 sub-cluster** each, resulting in **7 different sub-cluster**

	Lower ARPU (< IDR 150k)	Middle ARPU (IDR 150k – IDR 250k)	Top ARPU (>IDR 250k)
Low Transaction	**39.3k subs** IDR 86k 14 GB	**180 subs** IDR 166k 16 GB	
Potential Customer	**2139 subs** IDR 133k 20 GB	**12.4k subs** IDR 168k 25 GB	**108 subs** IDR 296k 25 GB
Data Addict		**1314 subs** IDR 232k 38 GB	**2078 subs** IDR 323k 46 GB



			Mean Value							
	Subsegment ARPU	Count MSISDN	total revenue	rev broadband	rev data pack	payload	los	first rank category SocialNet	first rank category Video	first rank category Transportation
Low Transaction	< 150K	39,319	85,789	80,012	77,951	14,061,724	1,281	5.46%	5.64%	48.79%
	Between 150K & 250K	180	166,440	80,319	77,467	16,102,707	1,145	2.78%	4.44%	50.56%
	> 250 K	-								
Potential Customer	< 150K	2,139	133,114	129,820	125,287	19,970,121	925	5.70%	6.08%	49.84%
	Between 150K & 250K	12,426	168,043	160,987	155,783	24,816,489	1,163	9.30%	8.14%	40.60%
	> 250 K	108	296,532	130,192	123,034	25,007,984	1,336	1.85%	13.89%	35.19%
Data Addict	< 150K	-								
	Between 150K & 250K	1,314	231,757	227,712	222,487	38,235,391	1,039	18.57%	12.86%	29.91%
	> 250 K	2,078	322,718	292,692	280,266	46,038,016	986	19.30%	16.89%	26.13%

DEPLOYMENT

GO-JEK berkontribusi Rp 8,2 triliun per tahun ke dalam perekonomian Indonesia melalui penghasilan Mitra Pengemudi.

		Sebelum menjadi mitra			Setelah menjadi mitra			Total Pendapatan yang masuk dalam perekonomian per bulan (Ribu Rupiah)		
Penghasilan Sebelum menjadi Mitra	Nilai Tengah (Ribu Rp)	Jumlah Responden (Survei)	Proporsi	Jumlah Responden *Weighted***	Jumlah Responden (Survei)	Proporsi	Jumlah Responden *Weighted***	Total Sebelum	Total Sesudah	Selisih
<1 juta	500	133	4%	27,081	39	1%	7,941	13,540,723.98	3,970,588.24	
1-1,5 juta	1250	302	9%	61,493	203	6%	41,335	76,866,515.84	51,668,552.04	
1,5-2 juta	1,750	707	21%	143,959	296	9%	60,271	251,928,733.03	105,475,113.12	
2-2,5 juta	2,250	982	30%	199,955	475	14%	96,719	449,898,190.05	217,618,778.28	
2,5-3,5 juta	3,000	799	24%	162,692	1148	35%	233,756	488,076,923.08	701,266,968.33	
3,5-6 juta	4,750	213	6%	43,371	1041	31%	211,968	206,012,443.44	1,006,849,547.51	
>6 juta	6,500	51	2%	10,385	113	3%	23,009	67,500,000.00	149,558,823.53	
Tidak Bekerja Sebelumnya	-	128	4%	26,063	0	0	-	-	-	
Total		**3315**	**100%**	**675,000***	**3315**	**100%**	**675,000**	**1,553,823,529.41**	**2,236,408,371.04**	**682,584,841.63**

*http://tekno.kompas.com/read/2017/12/18/07092867/berapa-jumlah-pengguna-dan-pengemudi-GO-JEK

**Weight berasal dari hasil survei yang telah diolah