Departemen Statistika
IPB University
StarCore analytics

SAS Data Science Competition

Memprediksi Keputusan Pengambilan Cashback Menggunakan Decision Tree

Biodata

Nama : Ravelto Wangistu
Usia : 22 tahun
Pekerjaan : Mahasiswa
Institusi : Universitas Indonesia
Domisili : Tangerang
Ketertarikan :

- Text mining
- Customer Analytics
- Machine Learning

Riwayat projek :

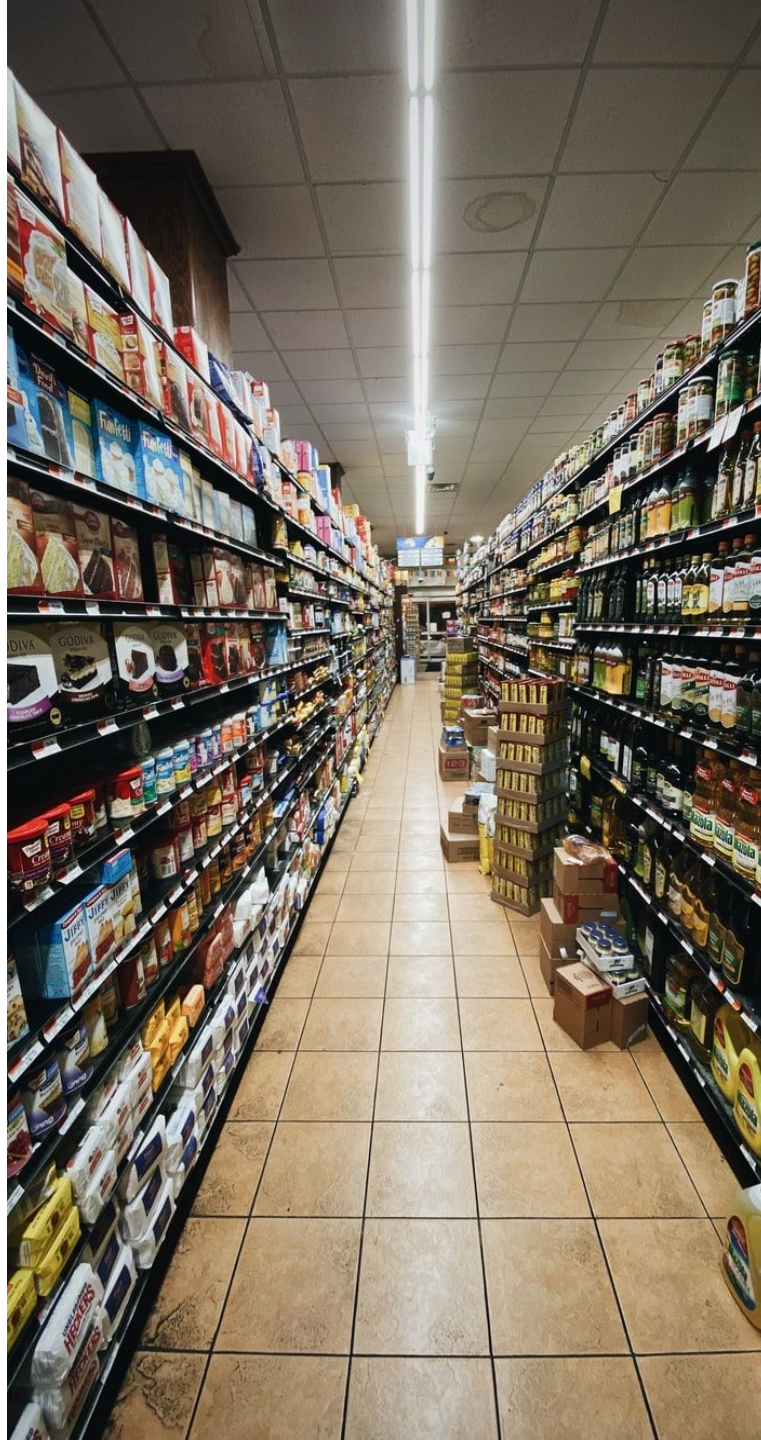
- Predictive modelling data Titanic Kaggle (link)
- Understanding Movie Datasets by SQL and Python (LINK)
- Exploratory Data Analysis in Hotel Business Datasets (LINK)

Departemen Statistika
IPB University
StarCore analytics

Daftar Isi

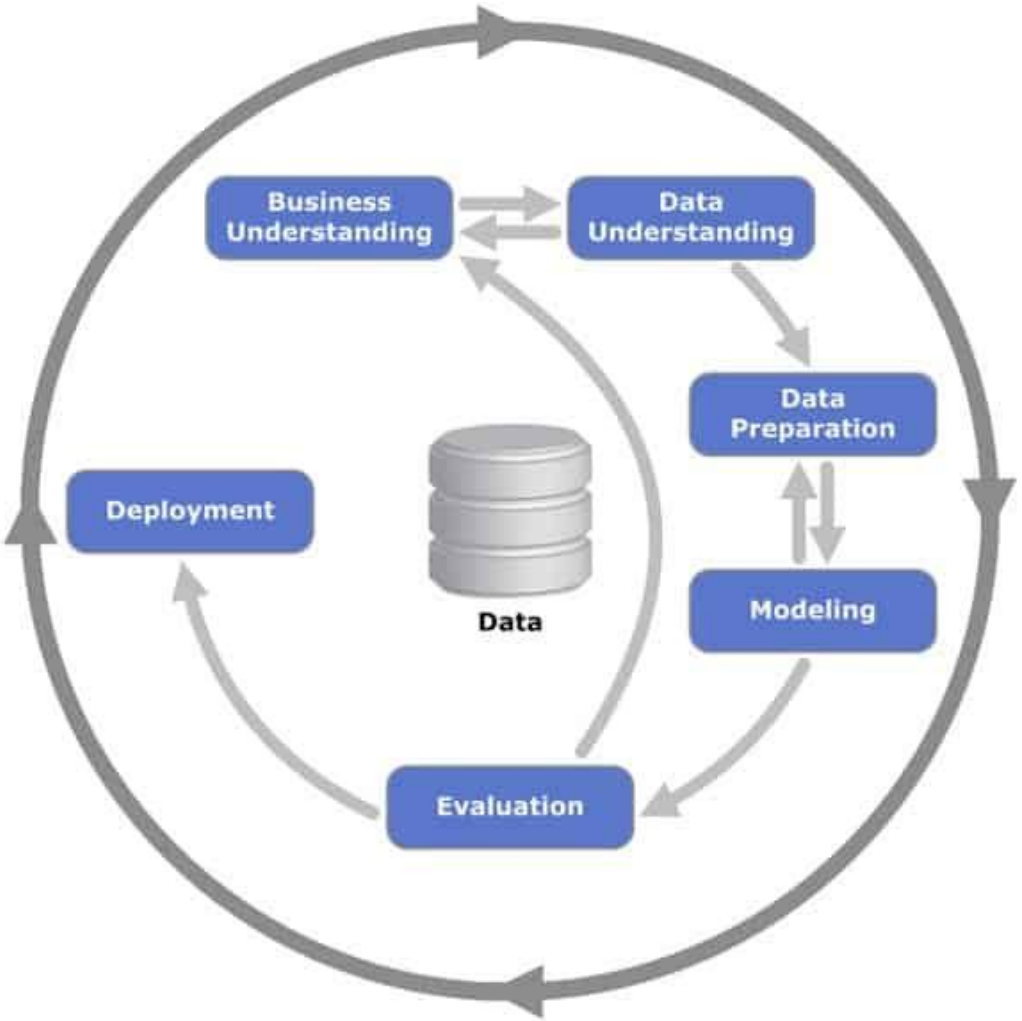
Departemen Statistika
IPB University

StarCore analytics



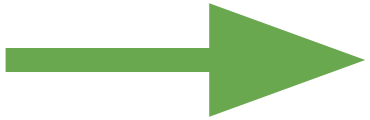
Presentasi ini akan menggunakan alur **CRISP-DM.**

CRISP-DM merupakan metode standarisasi yang dikeluarkan EU untuk data mining. Terdiri dari **Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, dan Deployment.**



Departemen Statistika
IPB University
StarCore analytics

BUSINESS UNDERSTANDING



Supermarket SAS-MART :

1. Promo cashback Rp 50K, setiap melakukan pembayaran sebesar Rp 800 K.
1. Promo dikirim ke 5000 customer.

Hasil:

Akan digunakan sebagai **penentuan member** mana yang akan diberikan promo.

Goal

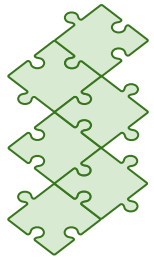
Tujuan	Output
Menentukan ***faktor-faktor apa yang mempengaruhi*** individu untuk menggunakan kupon tersebut (response).	**Model Machine Learning** berbentuk Decision Tree.
Melakukan ***evaluasi terhadap program promo cashback***, sekaligus mengembangkan promo supaya menjadi lebih baik.	**Mencari Insight** untuk mengevaluasi program cashback tersebut.

Departemen Statistika
IPB University
StarCore analytics

Departemen Statistika
IPB University
StarCore analytics

DATA UNDERSTANDING

Dari data yang didapat response_retail_v2.xlsx , beberapa fakta yang kita temukan adalah:



Data Shape

5000 baris &
20 kolumn

Data Null

Tidak ada data null dalam dataset ini.

Data Info

3 kolom (object) &
17 kolom (int)

```
df.shape
```

```
(5000, 20)
```

```
df.isnull().sum().max()
```

```
0
```

```
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   member_id             5000 non-null   int64
 1   gender                5000 non-null   object
 2   visit_last_1mo        5000 non-null   int64
 3   visit_last_2mo        5000 non-null   int64
 4   visit_last_3mo        5000 non-null   int64
 5   spending_last_1mo     5000 non-null   int64
 6   spending_last_2mo     5000 non-null   int64
 7   spending_last_3mo     5000 non-null   int64
 8   age                   5000 non-null   int64
 9   monthly_income        5000 non-null   int64
 10  marital_status        5000 non-null   object
 11  payment_channel       5000 non-null   object
 12  buy_groceries         5000 non-null   int64
 13  buy_toiletries        5000 non-null   int64
 14  buy_food              5000 non-null   int64
 15  buy_electronic        5000 non-null   int64
 16  buy_clothes           5000 non-null   int64
 17  buy_home_appliances   5000 non-null   int64
 18  recency_last_visit    5000 non-null   int64
 19  response              5000 non-null   int64
```

Keterangan dari setiap kolom yang ada dalam datasets ini :

Variable	Description
member_id	member's ID
gender	gender of member
visit_last_1mo	number of visits within month 1
visit_last_2mo	number of visits within month 2
visit_last_3mo	number of visits within month 3
spending_last_1mo	spending amount within month 1
spending_last_2mo	spending amount within month 2
spending_last_3mo	spending amount within month 3
age	age of members
monthly_income	member's monthly income

Variable	Description
marital_status	marital status of members
payment_channel	most frequent channel used
groceries	member buy groceries product in last 3 months (1: buy; 0: not buy)
toiletries	member buy toiletries product in last 3 months (1: buy; 0: not buy)
food	member buy food product in last 3 months (1: buy; 0: not buy)
electronic	member buy electronic product in last 3 months (1: buy; 0: not buy)
clothes	member buy clothes product in last 3 months (1: buy; 0: not buy)
home_appliances	member buy home appliances product in last 3 months (1: buy; 0: not buy)
recency_last_visit	number of days member last visit (recency)
response	dummy variable whether member taken the promo (1: yes; 0: no)

Categorizing Columns

Kolom tersebut dapat dibagi kedalam beberapa kategori yaitu :

Identifying Data

Member ID

Demographic Data

Gender, Age, Monthly_income, payment_channel, and Marital_Status

Product Data

groceries, toiletries, food, electronic, clothes, and home_appliances

RFM* Data

visit_last_1_mo, visit_last_2_mo, visit_last_3_mo, spending_last_1_mo, spending_last_2_mo, spending_last_3_mo, & recency_last_visit

**RFM stands for Recency, Frequency, Monetary*

Variables that we try to predict : Response (0 = No, 1 = Yes)

Departemen Statistika
IPB University
StarCore analytics

Software Yang Digunakan

Departemen Statistika
IPB University
StarCore analytics

DATA PREPARATION

Data Preparation : Changes

Untuk membantu menggali insight, penulis melakukan sedikit perubahan yaitu :

Membuat **Kolom Kumulatif Dari Total Visit dan Total Spend** yang dilakukan per bulan.

Dengan menggabungkan kolom kumulatif per bulan, maka kita bisa mengetahui total keseluruhannya, seperti :

Total Visit =
visit_last_1mo + visit_last_2mo + visit_last_3mo

Total Spending =
total_last_1mo + total_last_2mo + total_last_3mo

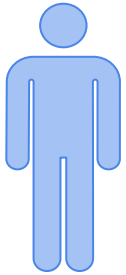
Melakukan Pembagian Kategori Terhadap Variable **Income dan Total Spend**

Untuk mendapatkan insight yang lebih mendalam, penulis melakukan pembagian kategori, seperti :

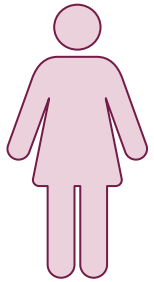
Monthly Income dibagi menjadi 3 kategori :
- Low Income (< Percentile 33.3)
- Medium Income (33.3% - 66.6%)
- High Income.

Total Spending dibagi menjadi 2 kategori :
- Low Spending (< nilai mean total spending)
- High Spending

Demographic



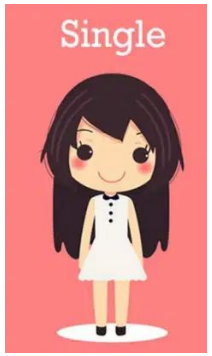
2002
(40.04%)



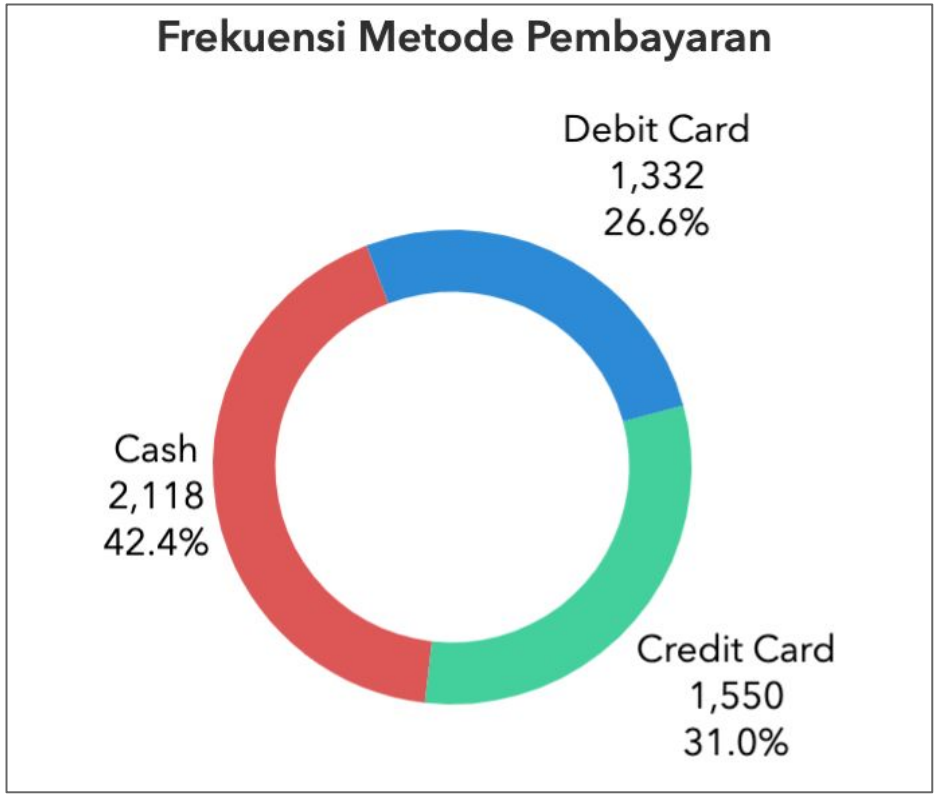
2998
(59.96%)

Age (Average)	Total Response	Total Spend (Average)
29.58	**2159** (43.18%)	**1,351,763**

3351
(67.02%)



1649
(32.98%)

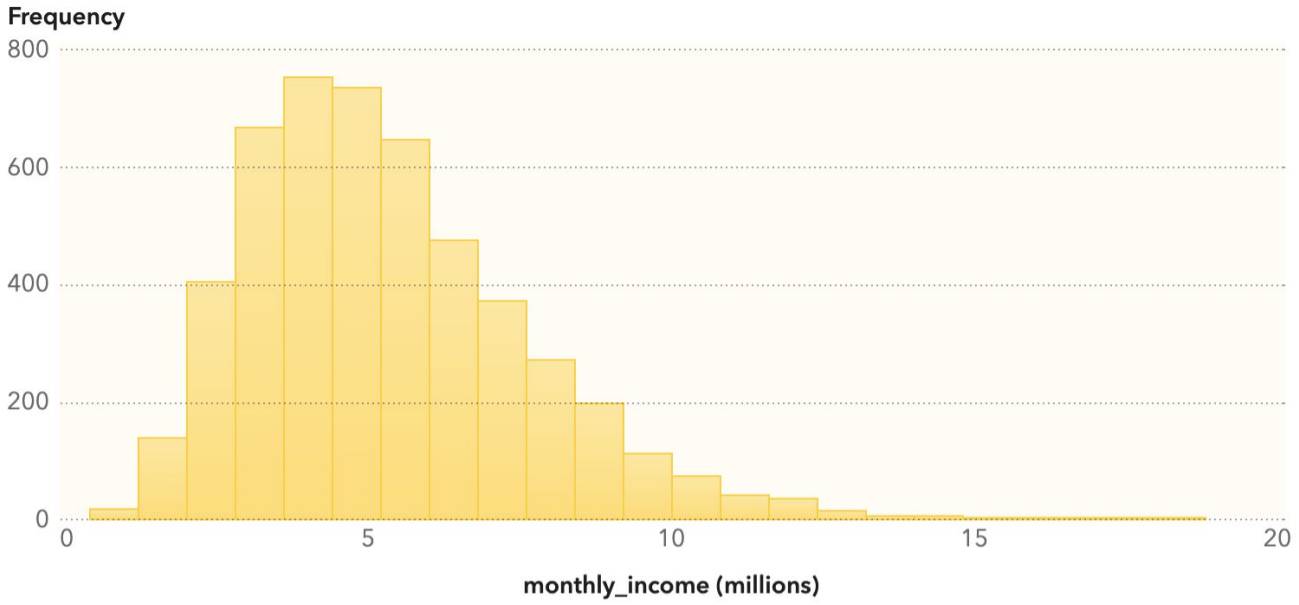


Departemen Statistika
IPB University
StarCore analytics

Demographic (2)

gender ▲	buy_clothes	buy_electronic	buy_food	buy_groceries	buy_home_appliances	buy_toiletries
Total	**480**	**397**	**4467**	**948**	**1215**	**3706**
Female	294	231	2669	565	709	2249
Male	186	166	1798	383	506	1457

Distribusi Monthly Income Berbentuk Positive Skew



Insight yang didapatkan :

a. Secara garis besar, individu yang datang ke SAS Mart adalah **perempuan, sudah menikah dan memiliki umur rata-rata 29.58 tahun.**

b. Kebanyakan individu melakukan pembayaran dengan **metode cash (42.4%).**

c. Lebih 50% individu yang berbelanja ke Sas Mart **membeli makanan (food)** dan ***toiletries.***

d. Total individu yang menggunakan **promo (response) sebesar 43.8%**

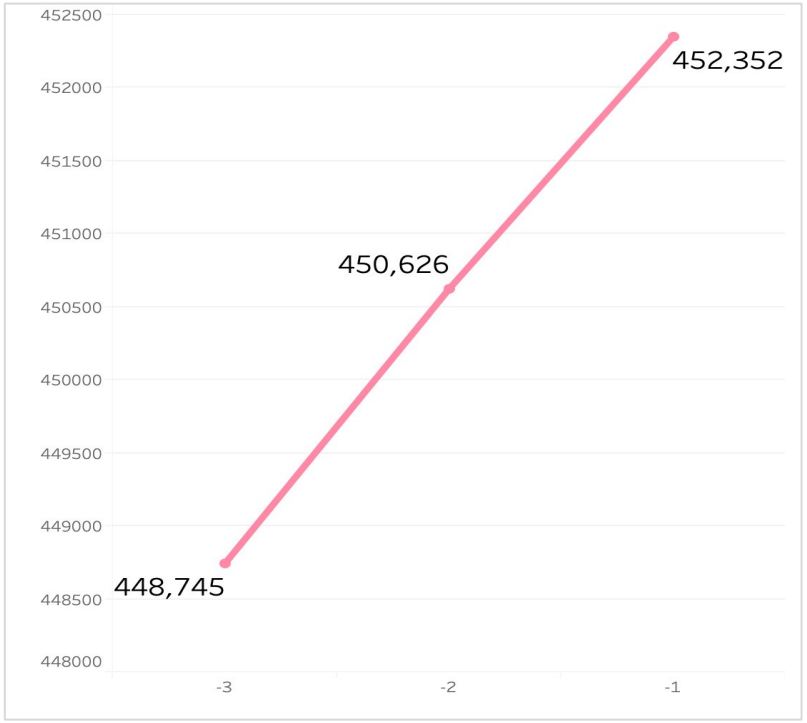
Departemen Statistika IPB University

StarCore analytics

Insights Yang Didapatkan

Terjadi peningkatan rata-rata spending setiap customer di setiap bulannya. Namun terdapat penurunan total visit yang signifikan pada satu bulan terakhir.

Average Total Spending Per Past Month



Months Ago

Average Total Visit Per Past Month

Months Ago

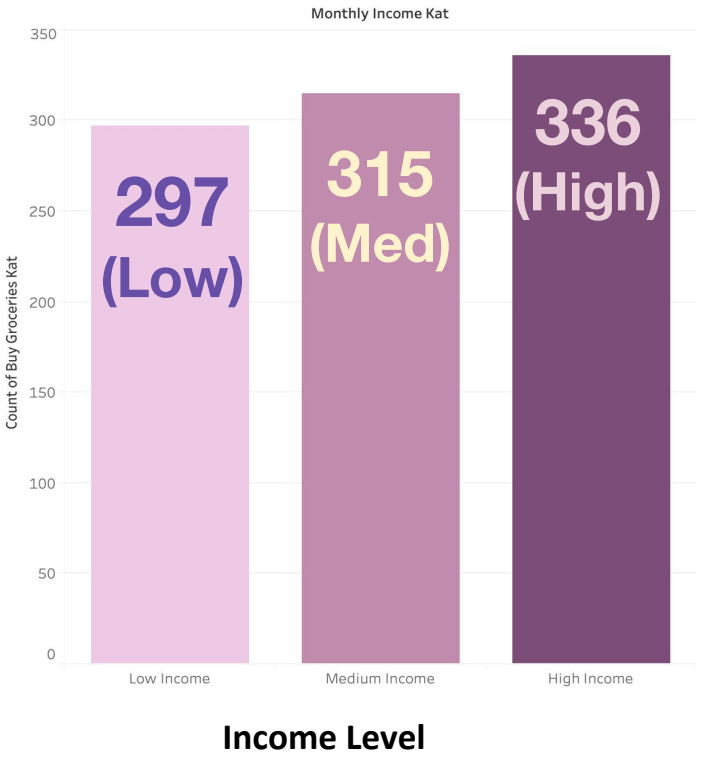
Departemen Statistika IPB University

StarCore analytics

Insights Yang Didapatkan (2)

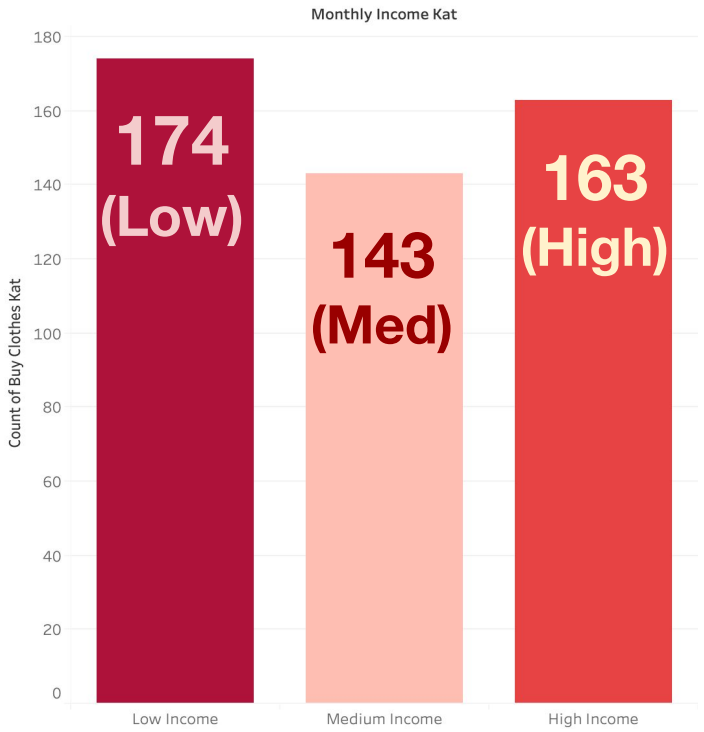
Pembelian ***Home Appliances*** dan ***Groceries*** didominasi oleh **High Income**,sedangkan Pembelian ***Clothes*** dan ***Electronic*** didominasi oleh **Low Income**. Pembelian ***Toiletries*** dan ***Food*** Hampir Merata.

Home Appliances (High Income)



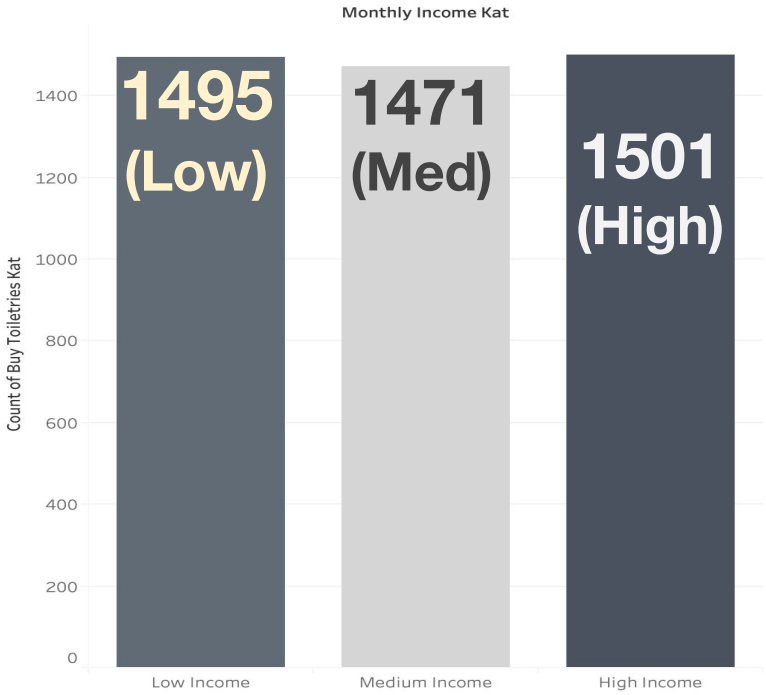
Income Level

Clothes (Low Income)



Income Level

Toiletries (Same)



Income Level

**) Low = Low Income, Med = Medium Income, & High = High Income*

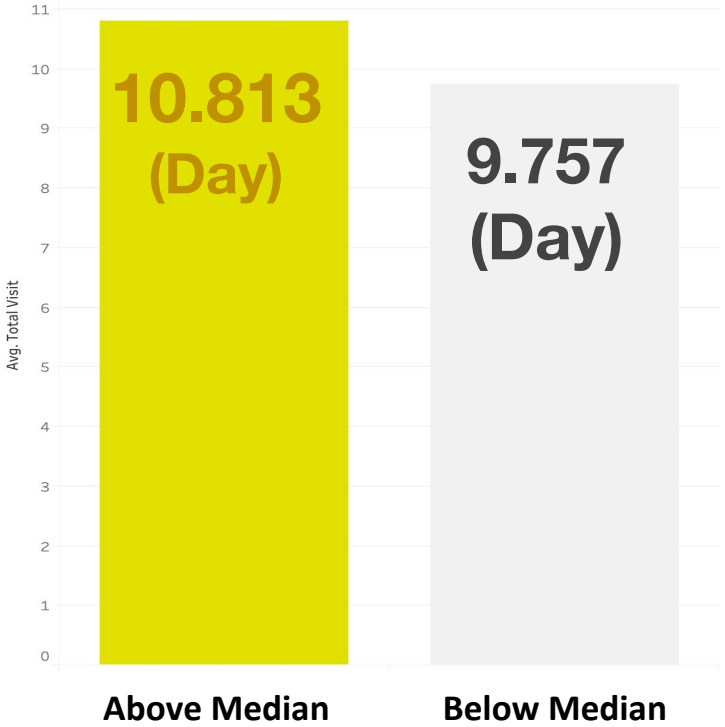
Departemen Statistika IPB University

StarCore analytics

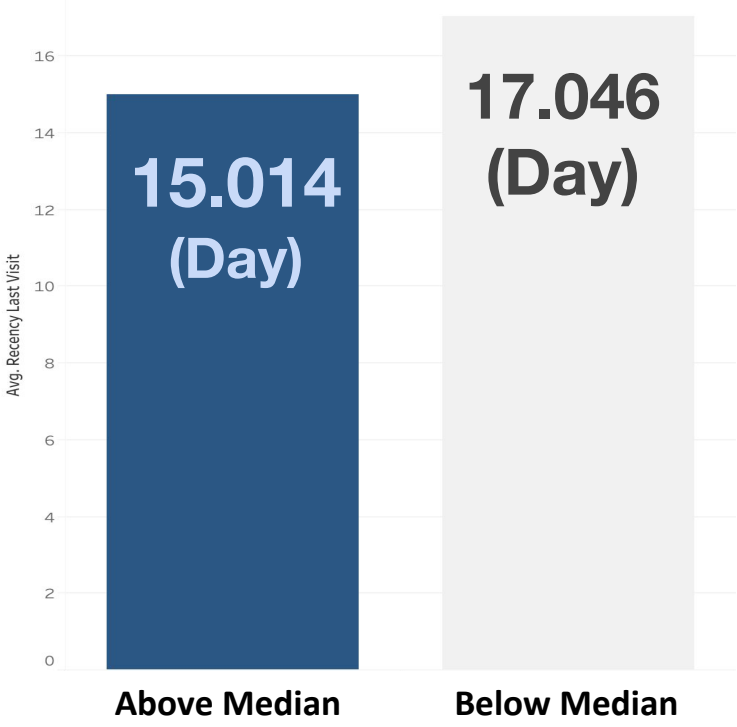
Insights Yang Didapatkan (3)

Total Spending Above Median memiliki ***average visit*** lebih banyak, ***recency rate*** lebih rendah, dan ***total response*** yang jauh lebih tinggi dari Spending Below Median.

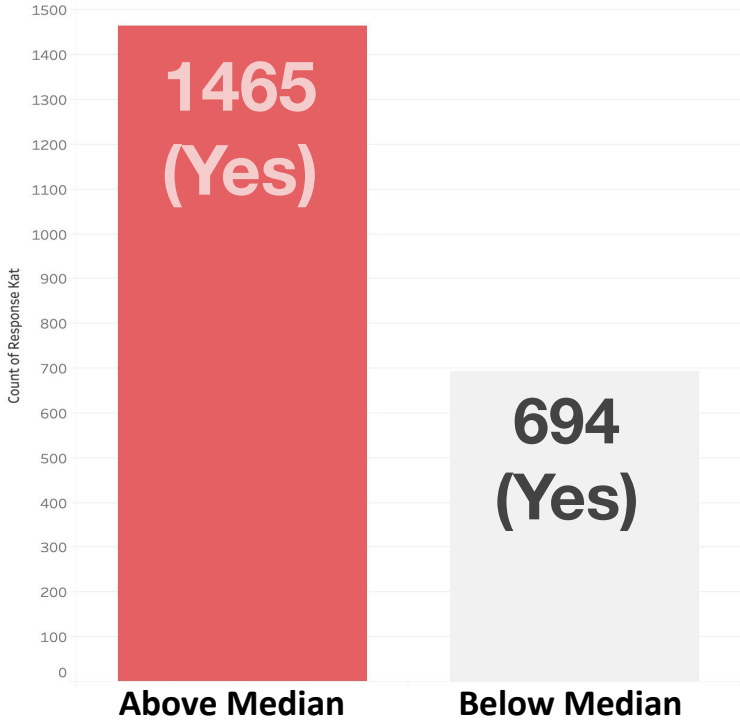
Total Visit Based on Spender



Total Recency Based on Spender



Total Response Based on Spender



*) *Low = Low Income, Med = Medium Income, & High = High Income*

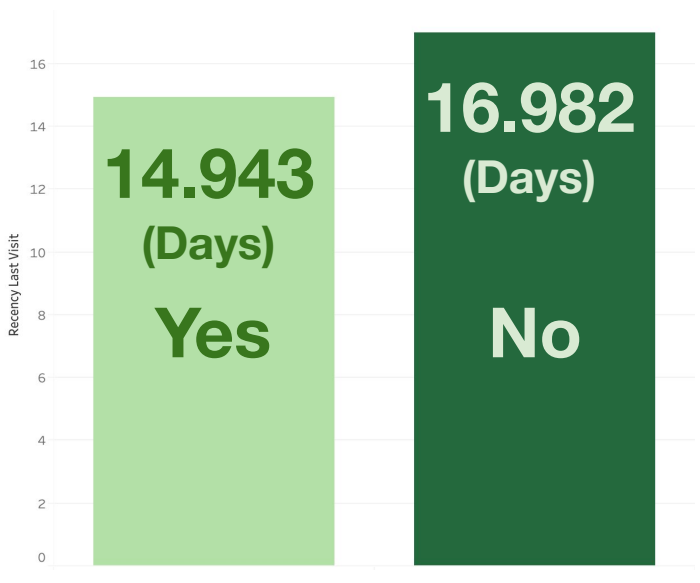
Departemen Statistika IPB University

StarCore analytics

Insight Yang Didapatkan (4)

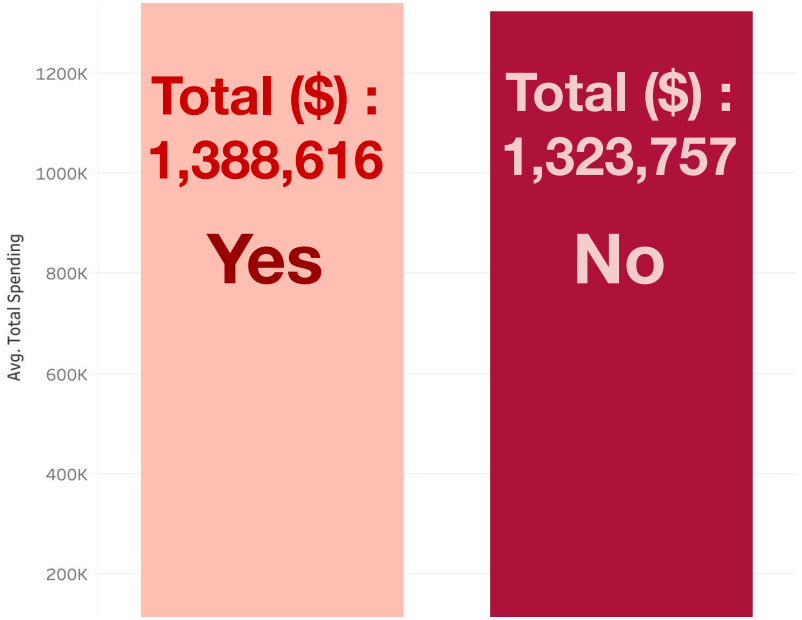
Response Rate sangat dipengaruhi oleh **Nilai Recency, Frequency, & Monetary Customers.** Semakin tinggi nilai RFM, semakin tinggi response ratenya.

Recency



Frequency

Monetary



Response Rate : (Yes / No)

Departemen Statistika IPB University

StarCore analytics

Departemen Statistika
IPB University
StarCore analytics

MODELLING

Persiapan Modelling : **Feature Importance**

Sebelum melakukan predictive analysis, ***feature importance*** dilakukan untuk mengurangi kompleksitas dan mencari ***10* feature*** yang paling berpengaruh berdasarkan *nilai effect size (r^2).*

Top 10 Feature Importance

	index	response
1	visit_last_2mo	0.02755094628
2	visit_last_3mo	0.02288436638
3	visit_last_1mo	0.02239169673
4	recency_last_visit	0.007696276876
5	spending_last_3mo	0.001694871532
6	spending_last_1mo	0.00150710662
7	spending_last_2mo	0.0014332801
8	buy_home_appliances	0.001237249966
9	monthly_income	0.0004480499849
10	member_id	0.0002770409238

*Feature peringkat 10 (member_id) tidak diambil karena member_id merupakan data identifying atau primary key.

Departemen Statistika IPB University

Modelling : **Find the best model**

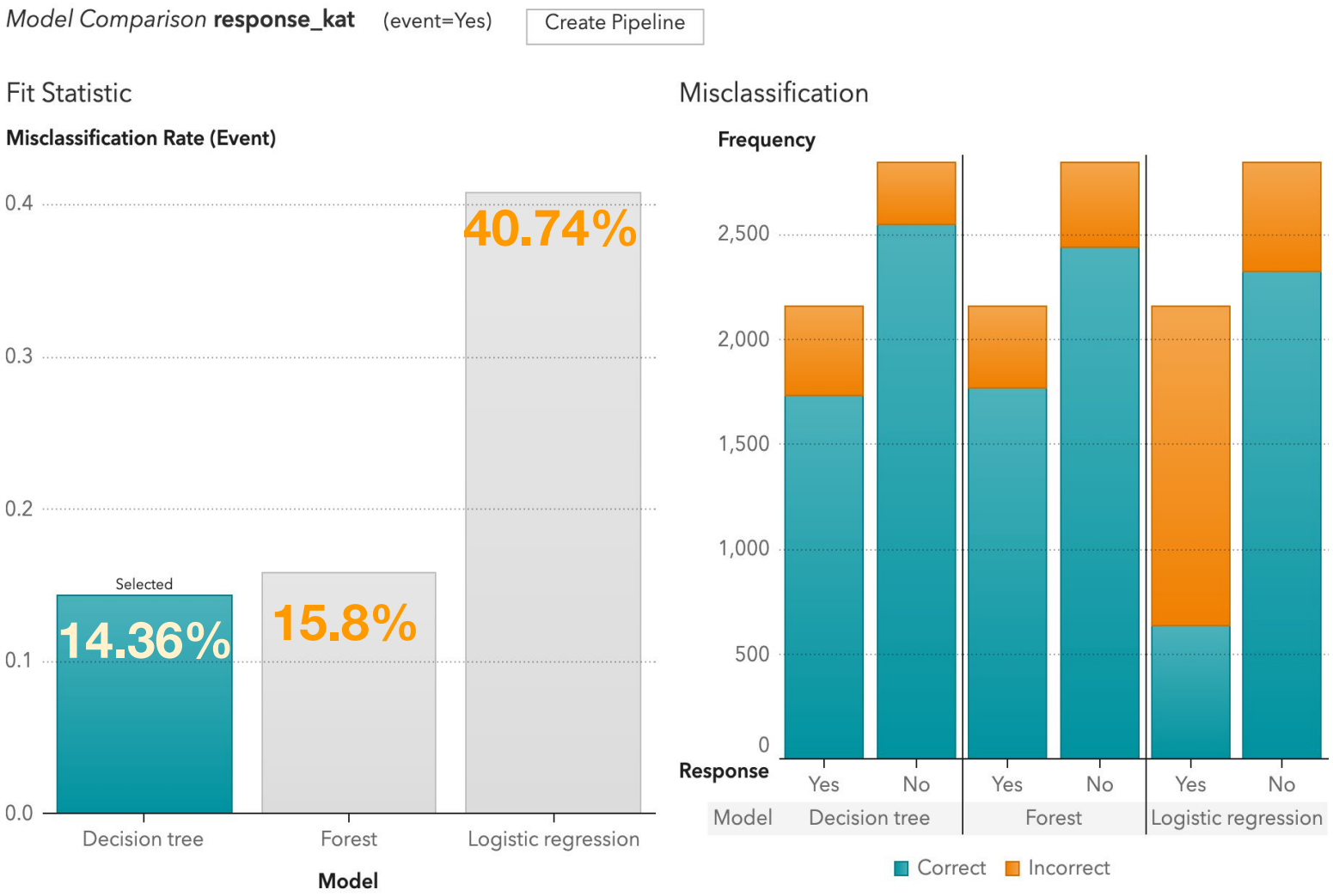
Berdasarkan fitur tersebut, penulis melakukan analisis dengan menggunakan 3 bentuk modelling yaitu, Logistic Regression**, Decision Tree,** dan Random Forest.

Tipe Modelling	**Alasan Digunakan / Tidak Digunakan**	**Link Appendix**
Decision Tree	Memiliki accuracy tinggi dan process model mudah dapat dimengerti.	LINK
Logistic Regression	Logistic Regression hanya cocok digunakan pada hubungan linear. Disisi lain, nilai false negative dari logistic regression sangatlah tinggi.	LINK
Random Forest	F1 Score dan misclassification rate random forest lebih tinggi daripada Decision Tree.	LINK

Penulis memutuskan untuk **memilih decision tree**, hal ini berdasarkan analisa model comparison yang dilakukan pada slide selanjutnya.

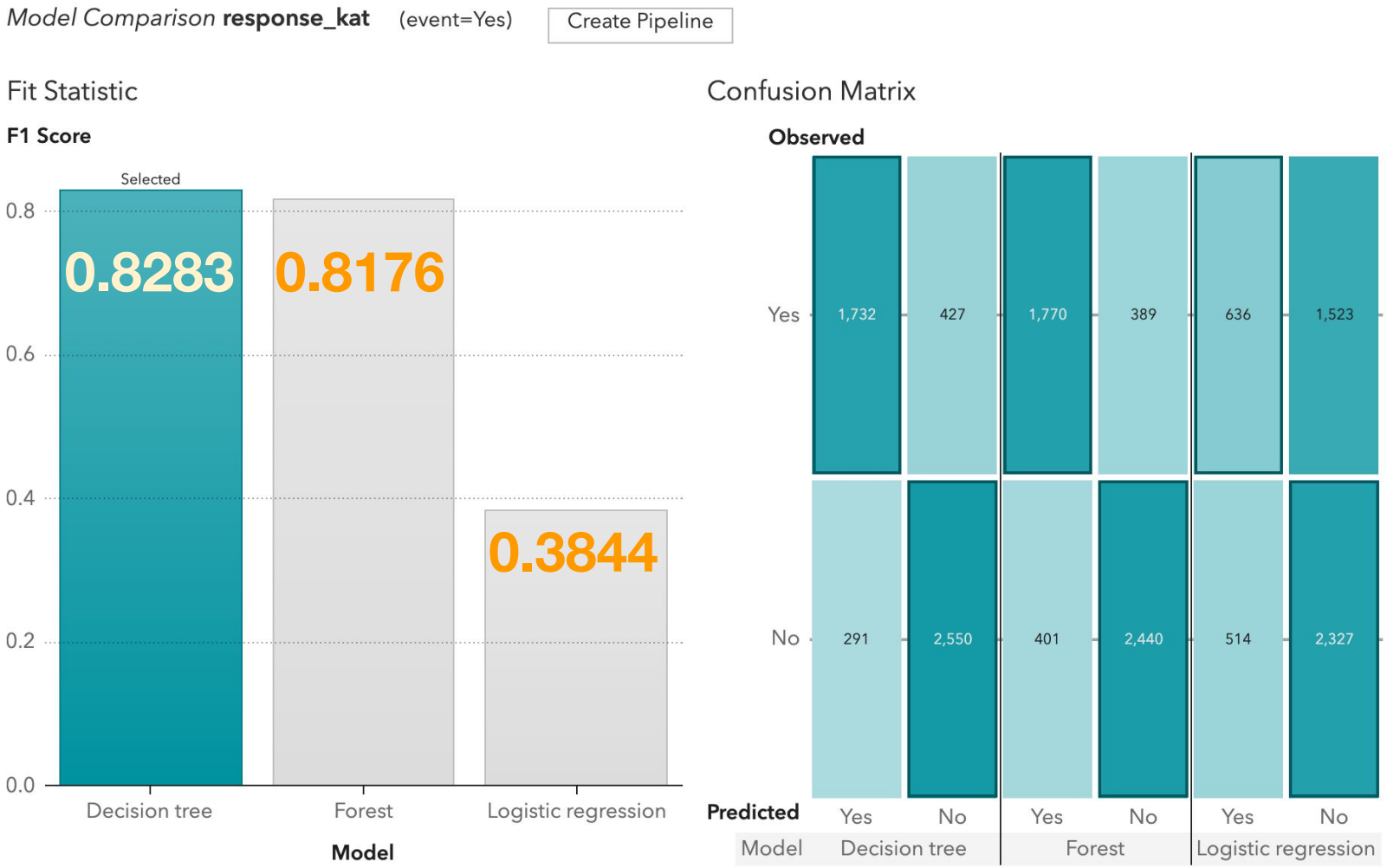
Model Comparison : Misclassification Rate

Decision tree memiliki **nilai misclassification rate yang paling rendah**. Sedangkan, logistic regression memiliki nilai tertinggi.

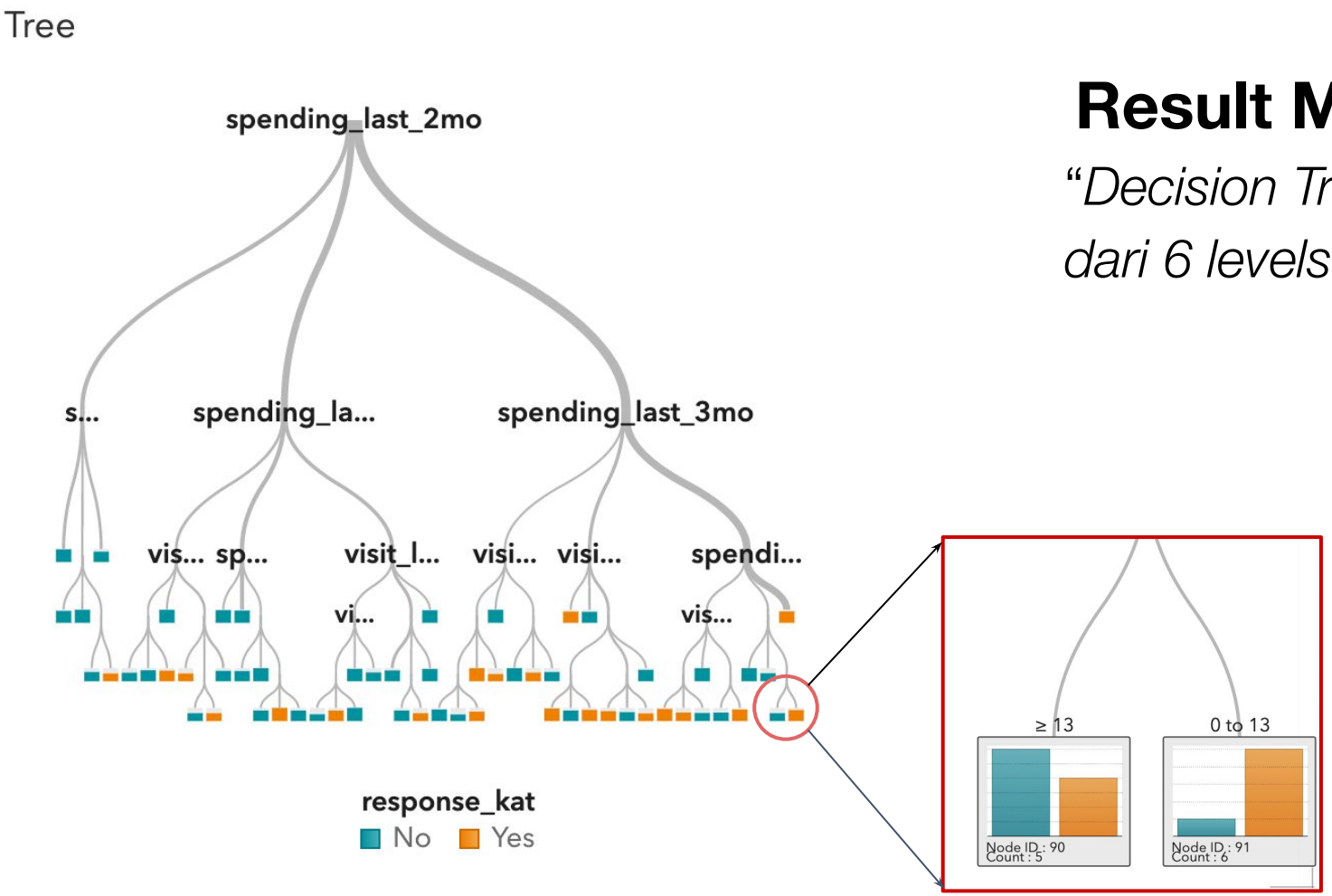


Model Comparison : F1 And Confusion Matrix

Berdasarkan nilai F1, **Decision Tree** (0.8283) memiliki nilai yang cukup tinggi. Selain itu, secara confusion matrix, nilai **decision tree dan random forest** memiliki true positive dan true negatif cukup baik.



Model : Decision Tree (2)



Result Model :

"Decision Tree yang terdiri dari 6 levels dan 3 branches."

Di dalam setiap nodes, terdapat **nilai probabilitas berapa** kemungkinan respon yang dihasilkan adalah "Yes" (mengambil cashback) atau "No" (tidak mengambil cashback).

Departemen Statistika
IPB University
StarCore analytics

Departemen Statistika
IPB University
StarCore analytics

EVALUATION

Evaluate : Variable Importance

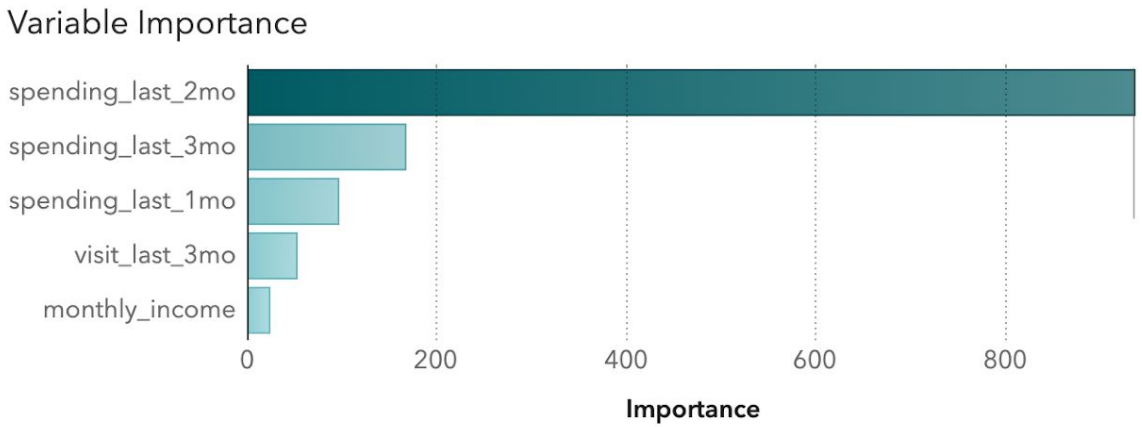


Table for 5th most feature importance :

Variable	Importance	Standard Deviation
spending_last_2mo	936.2311	312.119
spending_last_3mo	166.9813	78.7625
spending_last_1mo	96.1302	20.9085
visit_last_3mo	52.2731	12.9749
monthly_income	23.1381	1.2987

5 variabel yang memiliki pengaruh kuat terhadap decision tree yaitu:

- **spending_last_2mo**
- **spending_last_3mo**
- **spending_last_1mo**
- **visit_last_3mo**
- **monthly_income.**

Evaluate : Confusion Matrix

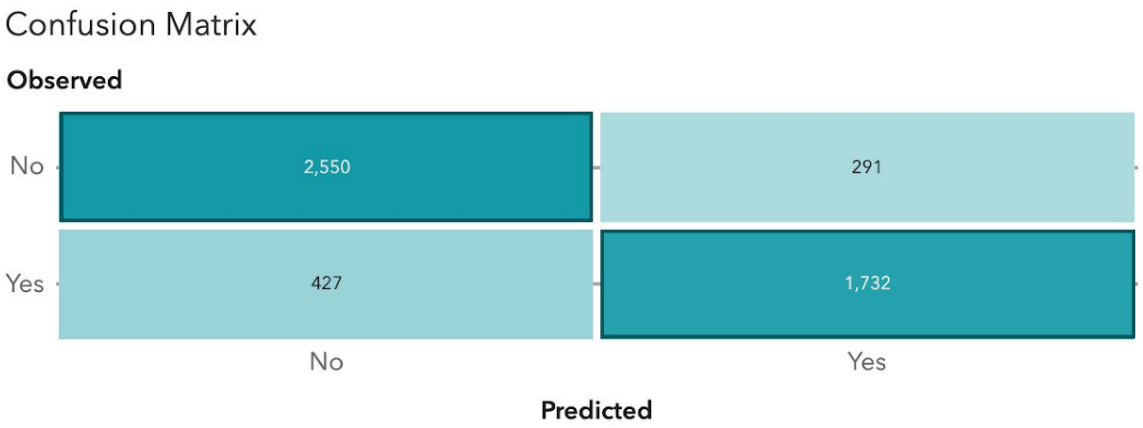


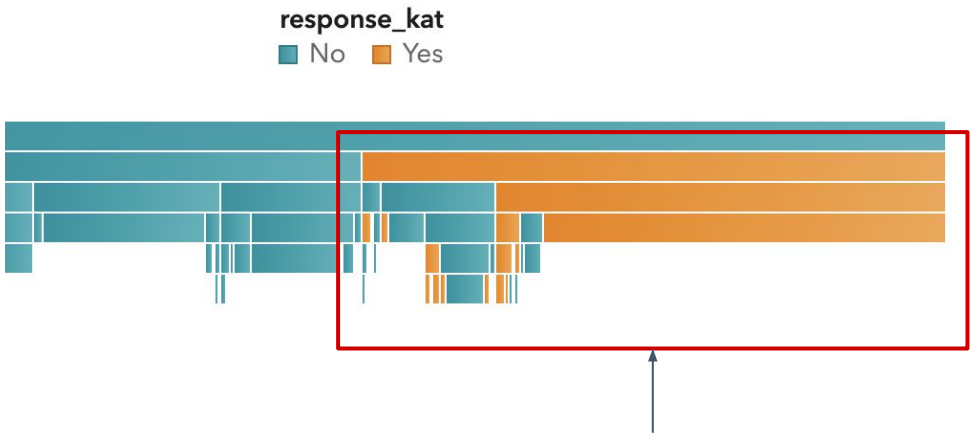
Table Confusion Matrix :

Status	Frequency	Percentage
True Negative	2550	89.76%
True Positive	1732	80.22%
False Negative	427	19.78%
False Positive	291	10.24%

Nilai true negative dan true positive dari model yang telah dibuat berada pada **level 80-90%**. Hal ini menunjukan bahwa **model memiliki nilai keakuratan yang cukup baik.**

Departemen Statistika
IPB University
StarCore analytics

Evaluate : Factor Leads To Response



Orange :
Kondisi dimana orang tersebut akan menggunakan cashback (Response Yes)

Tiga faktor yang paling dalam menentukan penggunaan cashback :

1. Melakukan spending pada 2 bulan lalu, sebesar **Rp 314.733 hingga Rp 691.928**. **(spend_last_2mo)**
1. Melakukan spending pada 3 bulan lalu, sebesar **Rp 356.391 hingga Rp 694.955. (spend_last_3mo)**
1. Melakukan spending pada 1 bulan lalu, sebesar **Rp 320.770 hingga Rp 657.167**. **(spend_last_1mo)**

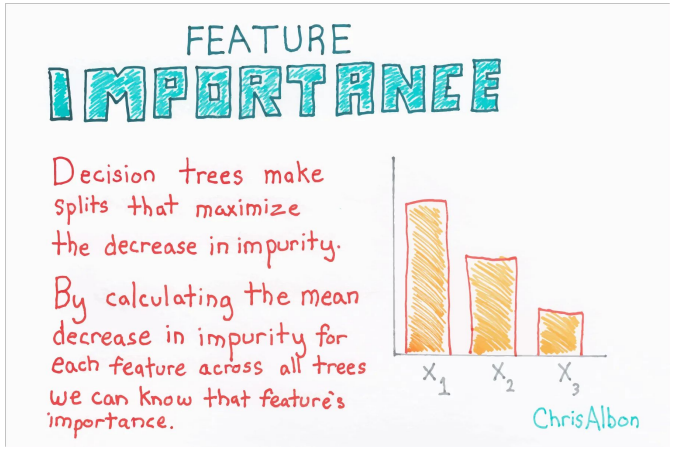
Individu yang memanfaatkan promo cashback, cenderung memiliki pengeluaran sebesar Rp. 300k-700k setiap bulannya

Departemen Statistika
IPB University
StarCore analytics

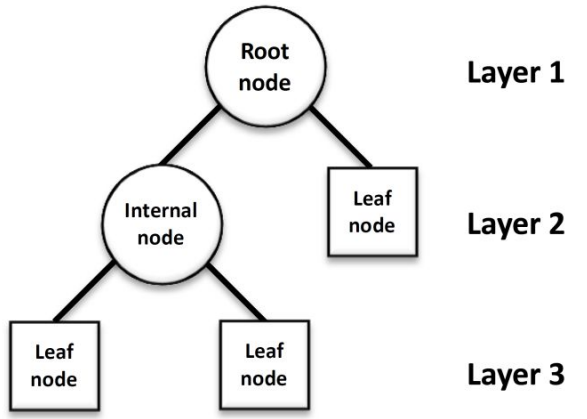
CONCLUSION

Conclusion : Goal 1

Goal 1 : Menentukan **faktor-faktor apa yang mempengaruhi** individu untuk menggunakan kupon tersebut (response).



Berdasarkan ***feature importance***, faktor yang memiliki pengaruh kuat dalam model adalah *data* ***RFM, monthly income, dan buy home appliances***.



Model Decision Tree digunakan karena nilai **misclassification rate and f1 score** yang lebih rendah daripada model lain.

Individu yang memanfaatkan promo *cashback* cenderung memiliki pengeluaran sebesar **Rp 300.000,00 - Rp 700.000,00** setiap bulannya

Departemen Statistika IPB University
StarCore analytics

Conclusion : Goal 2

Goal 2 : Melakukan ***evaluasi terhadap program promo cashback***, sekaligus mengembangkan promo supaya menjadi lebih baik.

No.	Problem	Solutions	Alasan
1.	Untuk mendapatkan cashback, **minimum pembelian sangatlah besar**.	Membuat **promo cashback lebih rendah** (contoh: Cashback 20 K untuk 200 K)	Rata-rata individu yang tertarik dengan promo response adalah kelompok individu yang **berbelanja di antara Rp 300.000 sampai Rp 700.000**
2.	Produk yang **penjualan tinggi** hanya **food dan toiletries**	Melakukan **Market Basket Analysis** untuk menentukan promosi yang spesifik kepada produk-produk yang kurang diminati.	Dengan membuat **promosi yang spesifik pada produk yang kurang diminati**, dapat **meningkatkan pembelian** produk tersebut.
3.	Pembayaran masih didominasi dengan **menggunakan cash.**	Memberikan **promo bagi individu yang melakukan pembayaran non-tunai** (seperti menggunakan kartu debit/kredit atau dompet-dompet digital).	Dengan memberikan promo, individu akan **lebih tertarik untuk menggunakan pembayaran non-tunai**, sekaligus meningkatkan kemungkinan **jumlah transaksi melewati pembelanjaan minimum** untuk mendapatkan cashback.

Departemen Statistika
IPB University

StarCore analytics

Terimakasih!

Departemen Statistika
IPB University
StarCore analytics

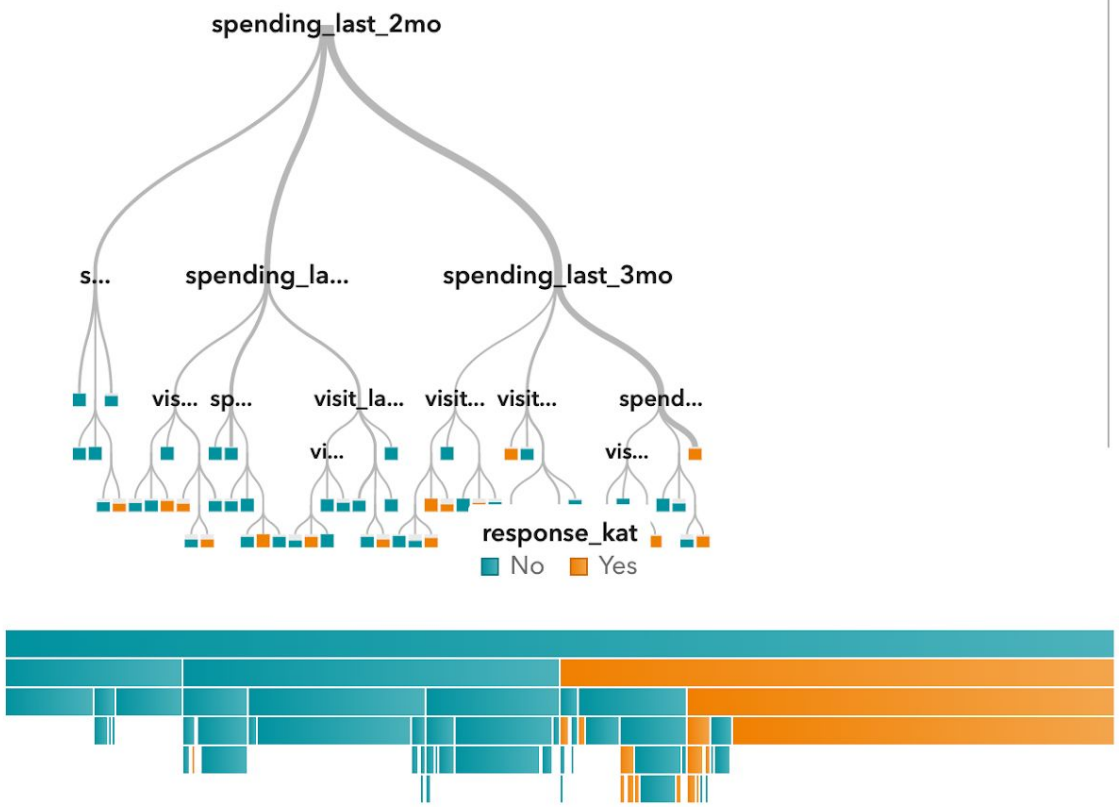
APPENDIX

Appendix : **Decision Tree**

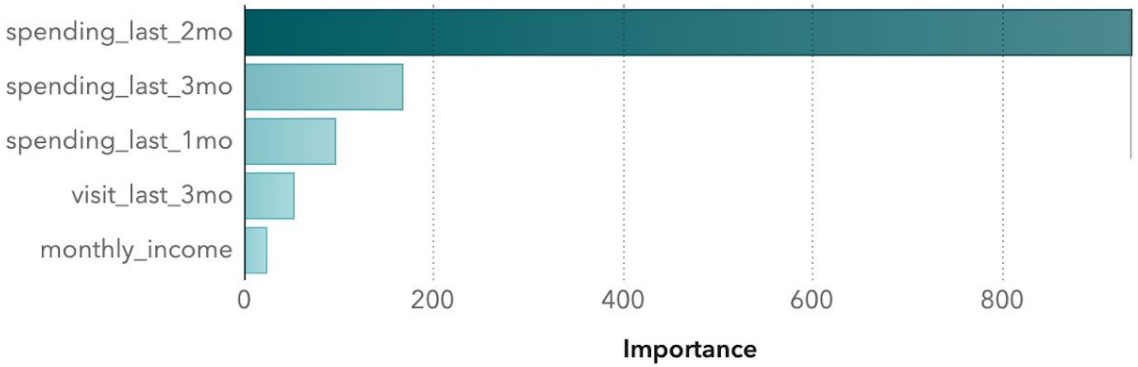
Decision Tree Retail

Decision Tree **response_kat** (event=Yes) F1 Score **0.828** Observations Used **5,000** Create Pipeline

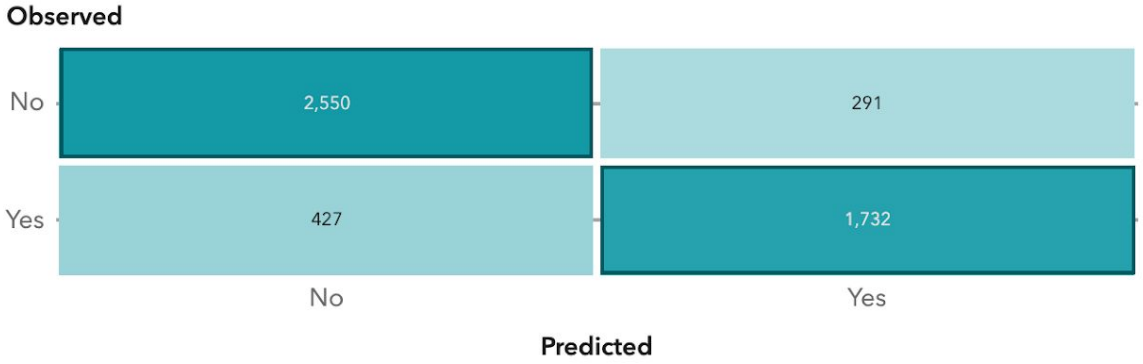
Tree



Variable Importance



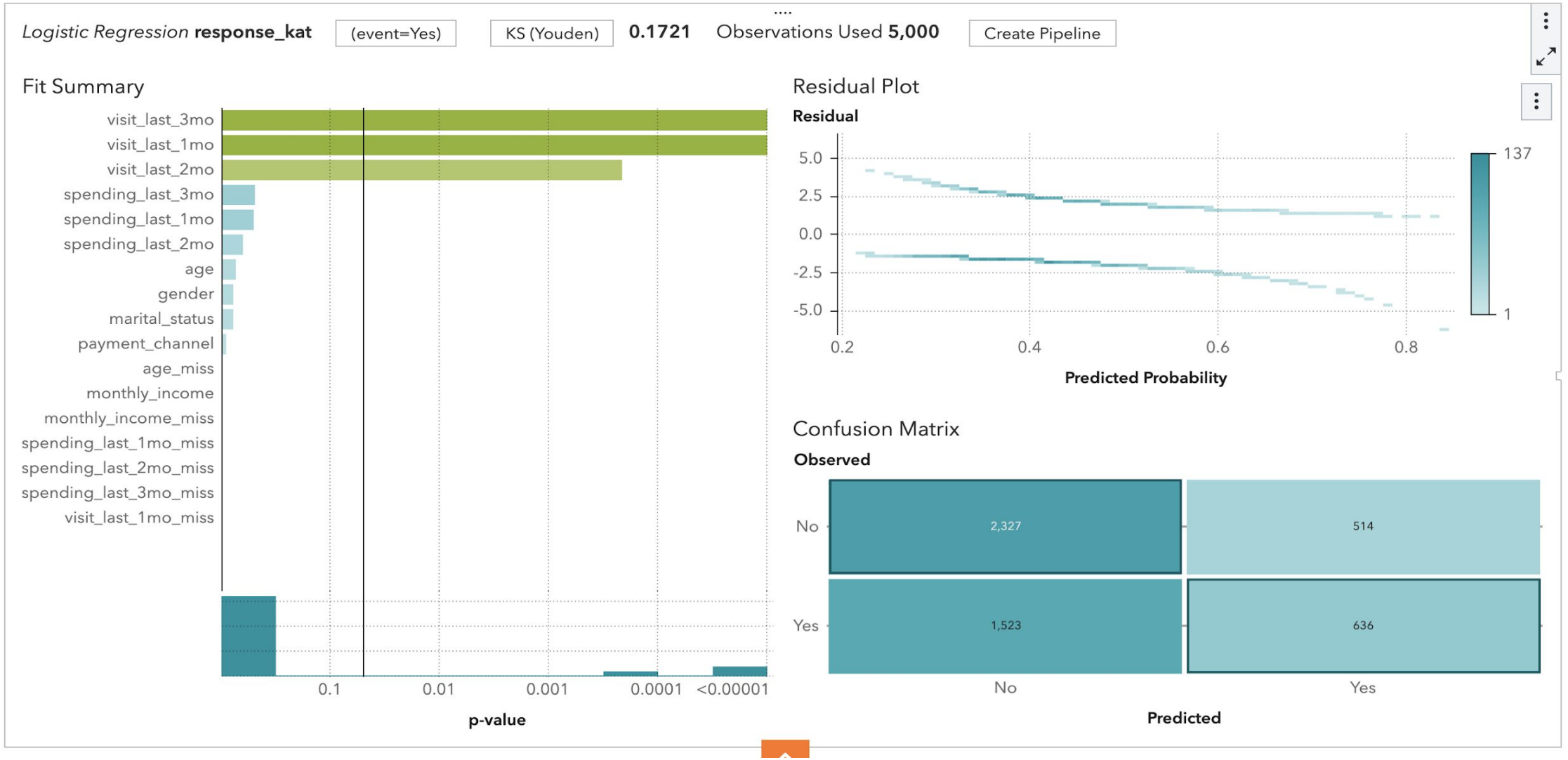
Confusion Matrix



Departemen Statistika IPB University

StarCore analytics

Appendix : **Logistic Regression**

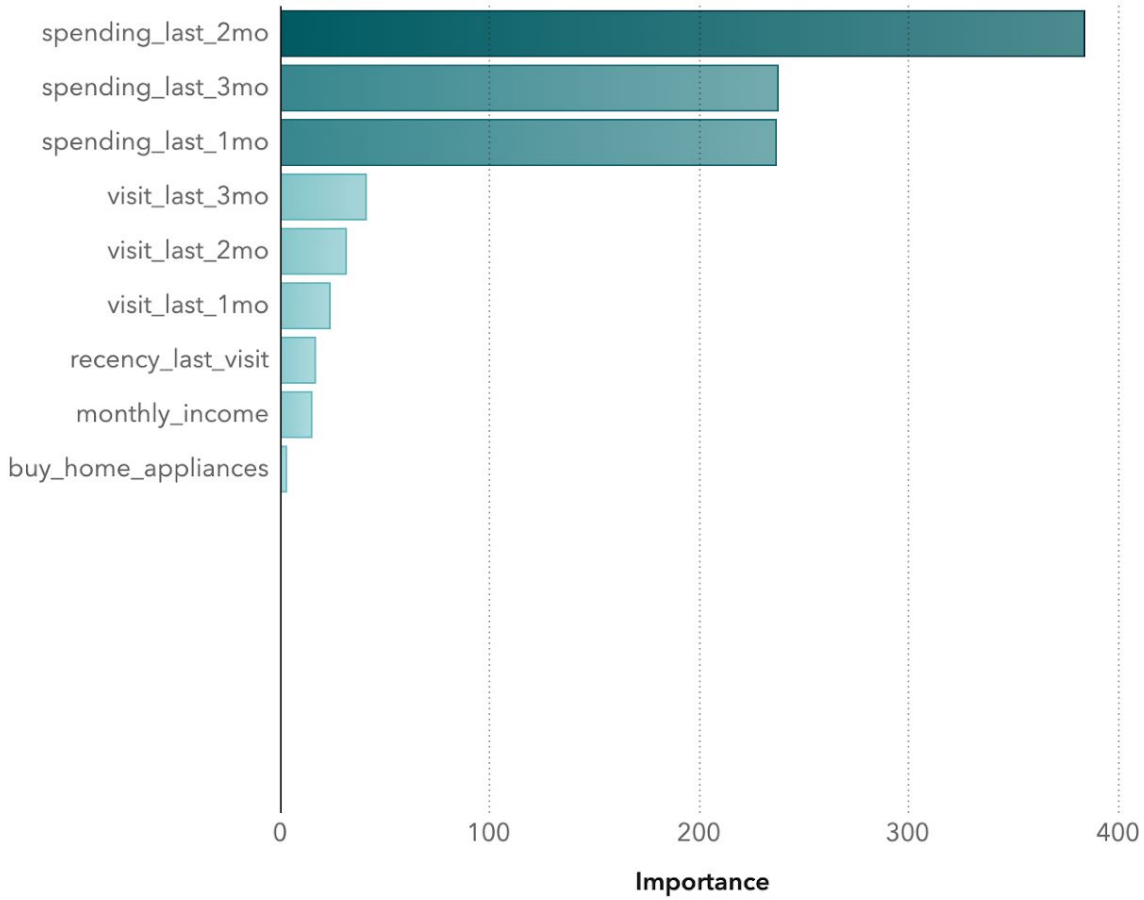


Departemen Statistika
IPB University
StarCore analytics

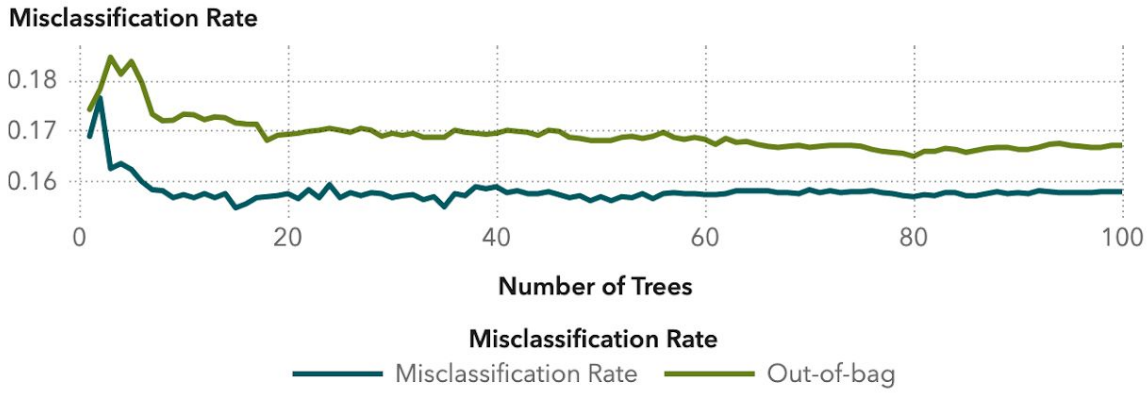
Appendix : **Random Forest**

Forest **response_kat** (event=Yes) KS (Youden) **0.6829** Observations Used **5,000** Create Pipeline

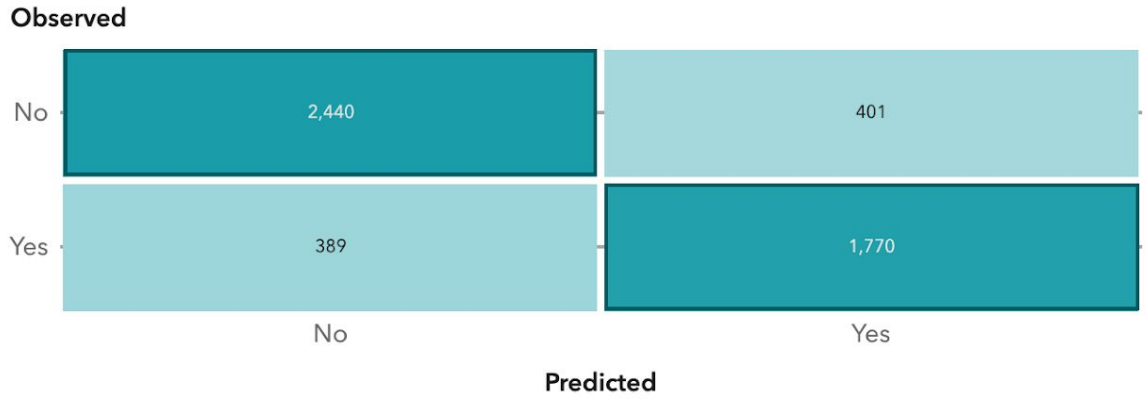
Variable Importance



Error Plot



Confusion Matrix



Departemen Statistika IPB University StarCore analytics