

# CAPSTONE 3 PROJECT

Travel Insurance

Disusun Oleh Reza Pradhana  
JCDS 0510 Jogja



# BUSINESS PROBLEM UNDERSTANDING



Suatu perusahaan yang bergerak di bidang asuransi perjalanan ingin mengetahui pemegang polis yang akan mengajukan klaim asuransi untuk pertanggungan. Data pemegang polis pada perusahaan asuransi merupakan data historis yang terdiri dari destinasi, produk asuransi, dan sebagainya.





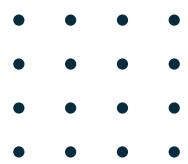
## TARGET

- 0 : Tidak mengajukan klaim asuransi
- 1 : Mengajukan klaim asuransi



## PROBLEM STATEMENT

- Perusahaan ingin mengetahui customer mana saja yang akan mengajukan klaim asuransi untuk pertanggungan





## GOALS

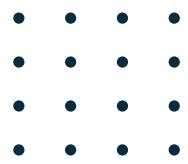
- Maka berdasarkan permasalahan tersebut, perusahaan ingin memiliki kemampuan untuk memprediksi customer yang akan mengajukan klaim asuransi atau tidak, sehingga dapat mempersiapkan dana yang dibutuhkan perusahaan tersebut.
- Dan juga perusahaan ingin mengetahui apa/faktor/variabel apa yang membuat customer biasanya melakukan klaim asuransi atau tidak, sehingga perusahaan dapat melakukan persiapan dana untuk pencairan asuransi.





## ANALYTIC APPROACH

- Jadi yang akan kita lakukan adalah menganalisis data untuk menemukan pola yang membedakan customer mana yang biasanya akan melakukan klaim asuransi dan yang tidak.
- Kemudian kita akan membangun model klasifikasi yang akan membantu perusahaan untuk dapat memprediksi probabilitas customer yang akan melakukan klaim asuransi atau tidak.

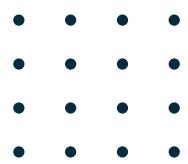




## METRIC EVALUATION

- Kita ingin membuat model yang dapat memprediksi customer mana yang akan melakukan claim asuransi.

Jadi nanti metric utama yang akan kita gunakan adalah roc\_auc





# ISI KOLOM

	Agency	Agency Type	Distribution Channel	Product Name	Gender	Duration	Destination	Net Sales	Commision (in value)	Age	Claim
0	C2B	Airlines	Online	Annual Silver Plan	F	365	SINGAPORE	216.0	54.00	57	No
1	EPX	Travel Agency	Online	Cancellation Plan	NaN	4	MALAYSIA	10.0	0.00	33	No
2	JZI	Airlines	Online	Basic Plan	M	19	INDIA	22.0	7.70	26	No
3	EPX	Travel Agency	Online	2 way Comprehensive Plan	NaN	20	UNITED STATES	112.0	0.00	59	No
4	C2B	Airlines	Online	Bronze Plan	M	8	SINGAPORE	16.0	4.00	28	No
...	...	...	...	...	...	...	...	...	...	...	...

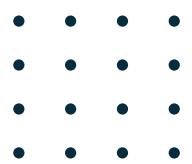
• • •  
• • •  
• • •  
• • •



# DATA CLEANING

---

- Mengatasi Missing Values yang ada.
- Menambahkan fitur gabungan



# MISSING VALUES

jumlah	
Gender	31647
Agency	0
Agency Type	0
Distribution Channel	0
Product Name	0
Duration	0
Destination	0
Net Sales	0
Commision (in value)	0
Age	0
Claim	0

Gender	jumlah	persentase %
	31647	71.39

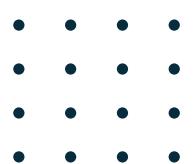


Handling

# MISSING VALUES

---

- Persentase missing value pada kolom gender terlalu besar yaitu 71.39% sehingga akan sulit untuk mengklasifikasikan pemegang polis yang melakukan klaim/tidak berdasarkan gender.



pengelompokan

# KOLOM AGE

---

- **Pada kolom Age terdapat error data, dimana age yang diinputkan ada yang melebihi 110 tahun terdapat sebanyak 676 data (kita asumsikan sebagai lansia)**  
**Kita akan membuat kolom baru bernama Group\_Age untuk mengelompokan usia customer**  
**Pengelompokan usia menurut Permenkes No 25. Tahun 2019 :**
  - **Usia Anak : 0 - 10 thn**
  - **Usia Remaja : 10 - 19 thn**
  - **Usia Dewasa : 19 - 44 thn**
  - **Usia Pra Lanjut Usia : 45 - 59 thn**
  - **Usia Lansia : >=60 thn**

...  
...  
...  
...



pengelompokan

# KOLOM DURATION

---

- agar mempermudah dalam melakukan analisa  
**Pengelompokan durasi travel sebagai berikut :**
  - **1-3 month : 0 - 90 days**
  - **4-6 month : 91 - 180 days**
  - **7-9 month : 181 - 270 days**
  - **10-12 month : 271 - 360 days**
  - **'>12 month : >361 days**

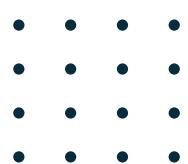


pembersihan

## KOLOM KOLOM NET SALES

---

- Terdapat 10 data Net Sales < 0, kemungkinan ini adalah kesalahan input Jadi akan kita hapus ke 10 data ini.



# DATA ANALYSIS

- 1. Customer yang melakukan Claim Asuransi paling banyak dari Agency C2B, meskipun secara persentase masih terbilang kecil.**
  - 2. Customer yang melakukan Claim Asuransi lebih banyak dari Airlines di banding Travel Agency**
  - 3. Customer yang melakukan Claim Asuransi lebih banyak menggunakan Distribution Channel Online di banding Offline, meskipun terpaut sedikit Produk Asuransi Annual Gold Plan, Annual Silver dan Annual Travel Protect Gold adalah 3 produk teratas yang paling sering dilakukan claim asuransi oleh customer**
- .....
- .....
- .....



# DATA ANALYSIS

**Group Age Remaja paling sering melakukan Claim Asuransi dibanding Group Age yang lain  
Customer yang memiliki Group Duration Travel >12 month paling sering melakukan claim dibanding Group Duration yang lain.**



# TUJUAN DESTINASI



Claim	No	Yes	count
<b>Destination</b>			
COSTA RICA	0.666667	0.333333	3
SINGAPORE	0.949384	0.050616	8199
CZECH REPUBLIC	0.953488	0.046512	43
ICELAND	0.958904	0.041096	73
ISRAEL	0.968750	0.031250	32
ITALY	0.979866	0.020134	298
TURKEY	0.980000	0.020000	50
FRANCE	0.981132	0.018868	318
SOUTH AFRICA	0.981308	0.018692	107

1. ada perbedaan kecenderungan claim antara tujuan destinasi negara yang berbeda.
2. Mari kita lihat kecenderungan 4 tujuan destinasi negara dengan jumlah customer terbanyak



# TUJUAN DESTINASI



Destination	count
SINGAPORE	8199
THAILAND	3658
MALAYSIA	3161
CHINA	2925
AUSTRALIA	2370
...	...

Terlihat bahwa destinasi MALAYSIA, THAILAND, dan CHINA memiliki kecenderungan customer yang sedikit melakukan claim asuransi, sedangkan untuk destinasi SINGAPORE lebih banyak customer yang melakukan claim asuransi.



# TAHAPAN PEMBUATAN MODEL

---

01

DATA PREPARATION

02

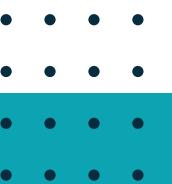
MODELING  
& EVALUATION

03

HYPERPARAMETER  
TUNING

04

SAVE MODEL



# DATA PREPARATION

---

Merubah fitur/kolom Agency menggunakan Binary Encoding, karena fitur ini memiliki unique data yang banyak dan tidak memiliki urutan/tidak ordinal, bila kita menggunakan One Hot Encoding akan terlalu banyak fitur yang terbuat, dan kalau kita menggunakan Ordinal/Label Encoding hasilnya dapat kurang cocok/ kurang baik. Oleh karena itu kita akan mencoba menggunakan Binary Encoding saja.

•  
•  
•  
•



**45%**  
Pengolahan Data



**85%**  
Evaluasi Hasil

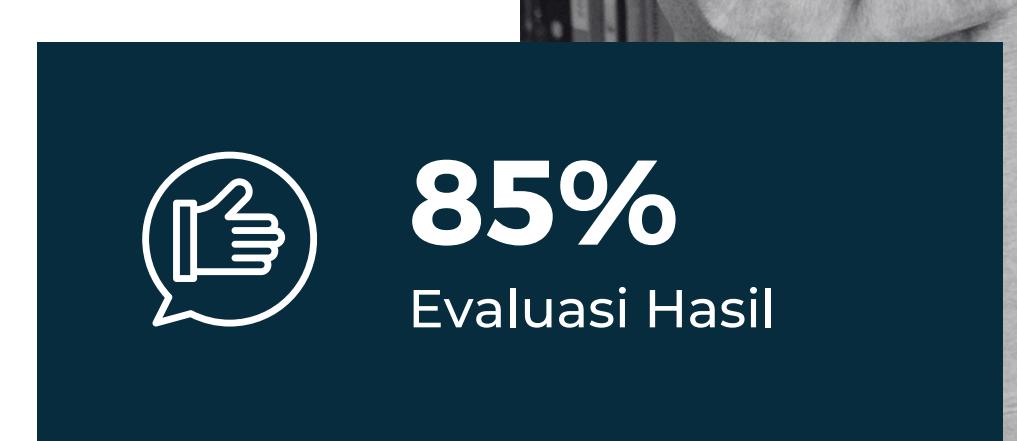


# DATA PREPARATION

---

Merubah fitur/kolom Agency Type menggunakan One Hot Encoding, karena fitur ini tidak memiliki urutan/tidak ordinal, dan juga jumlah unique datanya hanya sedikit.

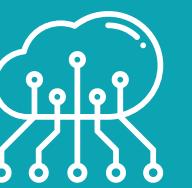
•  
•  
•



# DATA PREPARATION

Merubah fitur/kolom Distribution Channel menggunakan One Hot Encoding, karena fitur ini tidak memiliki urutan/tidak ordinal, dan juga jumlah unique datanya hanya sedikit.

•  
•  
•  
•  
•



**45%**

Pengolahan Data



**85%**

Evaluasi Hasil

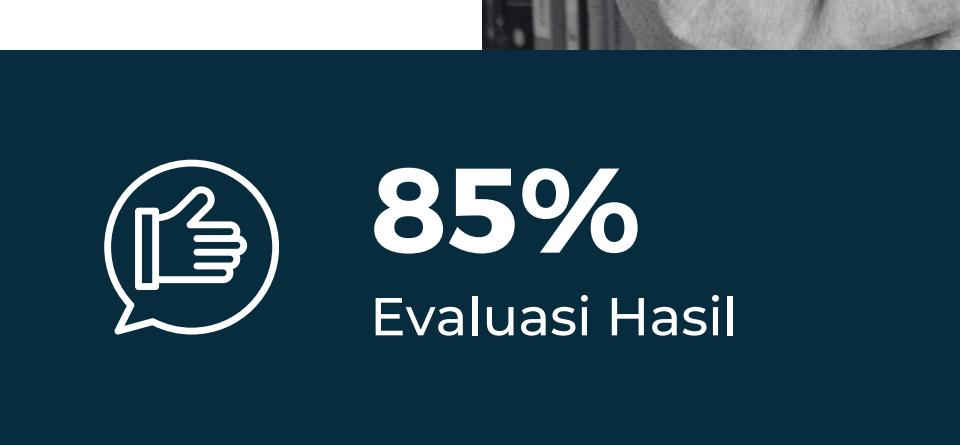


# DATA PREPARATION

---

Merubah fitur/kolom Product Name menggunakan Binary Encoding, karena fitur ini memiliki unique data yang banyak dan tidak memiliki urutan/tidak ordinal, bila kita menggunakan One Hot Encoding akan terlalu banyak fitur yang terbuat, dan kalau kita menggunakan Ordinal/Label Encoding hasilnya dapat kurang cocok/ kurang baik. Oleh karena itu kita akan mencoba menggunakan Binary Encoding saja.

• • •  
• • •  
• • •



# DATA PREPARATION

Merubah fitur/kolom Destination menggunakan Binary Encoding, karena fitur ini memiliki unique data yang banyak dan tidak memiliki urutan/tidak ordinal, bila kita menggunakan One Hot Encoding akan terlalu banyak fitur yang terbuat, dan kalau kita menggunakan Ordinal/Label Encoding hasilnya dapat kurang cocok/kurang baik. Oleh karena itu kita akan mencoba menggunakan Binary Encoding saja.



# DATA PREPARATION

Merubah fitur/kolom Group\_Age menjadi integer 0-4 dengan Ordinal Encoding, karena fitur ini adalah grouping usia customer, dimana terdapat group usia anak hingga lansia.

Merubah fitur/kolom Group\_Duration menjadi integer 0-4 dengan Ordinal Encoding, karena fitur ini adalah grouping durasi travel, dimana terdapat group durasi 1-3 month hingga lebih dari 12 month

⋮  
⋮  
⋮  
⋮  
⋮



# ENCODING

---



**Kita akan membagi dataset menjadi data training dan data test.  
Dimana 70% akan menjadi data training dan 30% sisanya akan menjadi data test**



# MODELING & EVALUATION



model	mean	roc_auc	sdev
CatBoost	0.791911	0.032410	
Logistic Regression	0.788468	0.042536	
LightGBM	0.780957	0.035623	
XGBoost	0.777270	0.030235	
GaussianNB	0.746376	0.040984	
Random Forest	0.687612	0.022446	
Decision Tree	0.651630	0.021224	
KNN	0.611717	0.021180	

Terlihat bahwa model CatBoost adalah yang terbaik untuk roc\_aucnya dari setiap model yang menggunakan default hyperparameter. LightGMB, XGBoost dan Logistic Regression juga memiliki hasil yang baik hampir sama dengan CatBoost



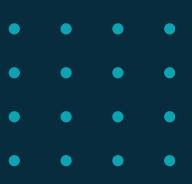
# PENGUJIAN K CROSS VALIDATION



**K-fold cross validation** adalah teknik untuk mengukur kinerja model secara menyeluruh, mencegah overfitting, dan memastikan generalisasi yang baik pada data yang belum terlihat. Proses ini melibatkan pembagian dataset menjadi "k" bagian yang sama besar.



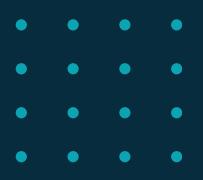
# EVALUATION METRICS WITH OVERSAMPLING



	Train Accuracy	Test Accuracy	Train ROC AUC	Test ROC AUC	Train F1 Score	Test F1 Score	Train Recall	Test Recall	Train Precision	Test Precision
0	0.917735	0.938001	0.961539	0.976629	0.917493	0.859776	0.971877	0.939068	0.876921	0.631325
1	0.918972	0.927425	0.962406	0.963818	0.918758	0.838824	0.970352	0.870307	0.879931	0.612981
2	0.916251	0.938001	0.961842	0.968087	0.916028	0.863133	0.967754	0.874598	0.877379	0.674938
3	0.917220	0.944566	0.962718	0.978131	0.917017	0.873056	0.966723	0.914089	0.879634	0.676845
4	0.916127	0.935084	0.961597	0.970082	0.915899	0.853897	0.968207	0.872483	0.876872	0.650000
5	0.916127	0.939096	0.962417	0.973365	0.915899	0.860001	0.968207	0.911348	0.876872	0.644110
6	0.917038	0.932896	0.962629	0.974332	0.916805	0.843172	0.969899	0.920152	0.877163	0.597531
7	0.916172	0.938366	0.963413	0.976183	0.915938	0.860084	0.968869	0.915789	0.876492	0.642857
8	0.919079	0.938366	0.962583	0.972018	0.918819	0.867848	0.975672	0.907643	0.876468	0.670588
9	0.914852	0.940168	0.961318	0.973406	0.914587	0.865710	0.970600	0.896321	0.873238	0.668329
Average	0.916957	0.937197	0.962246	0.972605	0.916724	0.858550	0.969816	0.902180	0.877097	0.646950



# EVALUATION METRICS WITH OVERSAMPLING



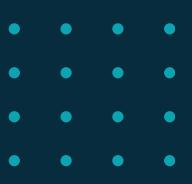
Terlihat bahwa hasil evaluation metrics untuk recall class positive sudah cukup baik saat di oversampling.

Noted\* Dimana sebelumnya saya sudah mencoba berbagai macam evaluation metrics seperti oversampling, undersampling, smote, borderline-smote, adasyn, without oversampling/undersampling. Hasil yang paling baik diperlihatkan pada saat oversampling.

Tidak semua saya tampilkan, untuk kepentingan user agar lebih mudah membaca hasil analisa, hanya yang terbaik yang saya tampilkan.



# CLASSIFICATION REPORTS WITH OVERSAMPLING



	precision	recall	f1-score	support
0	0.99	0.94	0.96	2463
1	0.62	0.93	0.75	279
accuracy			0.94	2742
macro avg	0.81	0.93	0.86	2742
weighted avg	0.95	0.94	0.94	2742

	precision	recall	f1-score	support
0	0.98	0.93	0.96	2449
1	0.61	0.87	0.72	293
accuracy			0.93	2742
macro avg	0.80	0.90	0.84	2742
weighted avg	0.94	0.93	0.93	2742

	precision	recall	f1-score	support
0	0.98	0.95	0.97	2431
1	0.70	0.86	0.77	311
accuracy			0.94	2742
macro avg	0.84	0.91	0.87	2742

Terlihat bahwa model yang setelah di oversampling memiliki recall dari kedua class yang cukup baik. Oleh karena itu untuk kasus kali ini, mari kita gunakan model yang menggunakan oversampling.



# CLASSIFICATION REPORTS

---

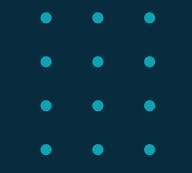


Classification Report Tuned CatBoost:				
	precision	recall	f1-score	support
0	0.98	1.00	0.99	11549
1	0.00	0.00	0.00	202
accuracy			0.98	11751
macro avg	0.49	0.50	0.50	11751
weighted avg	0.97	0.98	0.97	11751



# CONCLUSION & RECOMMENDATION

---



Hasil model sebelum tuning jauh lebih baik dibandingkan setelah tuning, terutama karena mampu menangani kelas minoritas dengan lebih baik. Setelah tuning, model menjadi bias terhadap kelas mayoritas, meskipun akurasi keseluruhan meningkat.

hal ini akan kita perbaiki di kesempatan berikutnya dengan penambahan kolom/fitur yang lebih berhubungan dengan customer melakukan claim asuransi atau tidak.



# CONCLUSION & RECOMMENDATION

---



## Recommendation

- Menambah banyaknya data customer yang melakukan claim asuransi
- Menambahkan fitur2 atau kolom2 baru yang kemungkinan bisa berhubungan dengan claim asuransi, seperti tanggal claim asuransi (untuk melihat ada atau tidak di bulan2 tertentu berkorelasi dengan banyaknya claim asuransi yang masuk) dan ID Customer (untuk melihat customer yang melakukan claim asuransi apakah new\_customer atau customer lama yang sudah berlangganan travel asuransi)
- Menganalisa data-data yang model kita masih salah tebak untuk mengetahui alasannya dan karakteristiknya bagaimana.



# TERIMA KASIH



Disusun Oleh Reza Pradhana  
JCDS 0510