Travel Insurance Claim Prediction Utilizing Machine Learning

Capstone Project Presentation – Machine Learning

Rafif Razaan Ananto

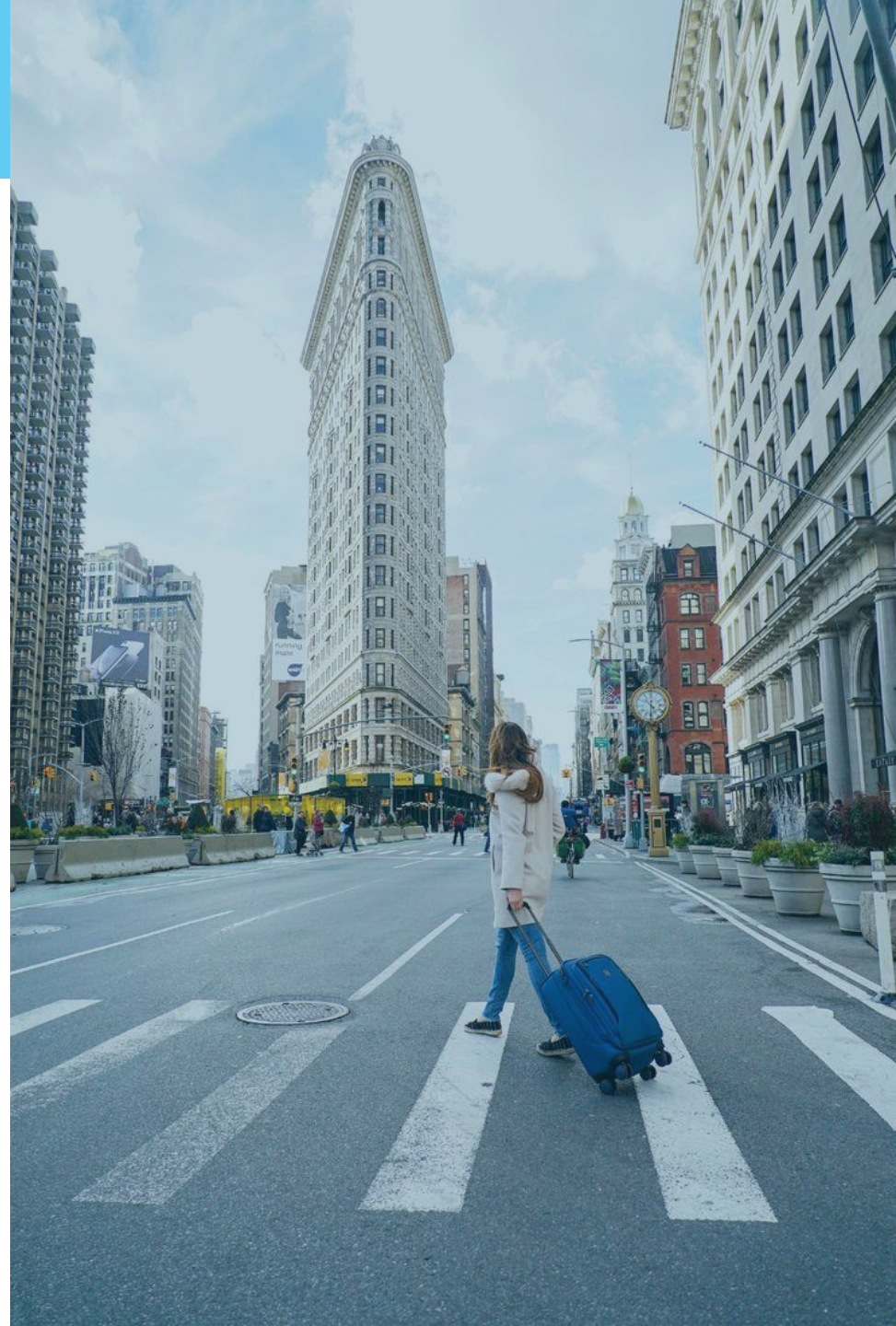


Table of Contents

Business Understanding

About Company

Perusahaan asuransi memiliki bisnis model risk pooling. Untuk mencari keuntungan, Perusahaan asuransi harus mencari nasabah/pemegang polis sebanyak-banyak nya

Line of business

Asuransi General

Asuransi Kesehatan

Asuransi Jiwa

Business Problem

Sumber utama pendapatan asuransi berasal dari premi, namun Perusahaan asuransi perlu menentukan premi yang tepat

Salah satu source cost Perusahaan asuransi adalah operational cost, salah satu nya adalah cost proses assessment claim

Objective and Goals

1. Mengetahui calon pemegang polis yang memiliki kecenderungan untuk claim
2. Meningkatkan pendapatan Perusahaan dengan menentukan premi yang sesuai
3. Mempermudah proses underwriting nasabah sehingga menghemat biaya operasional

Data Understanding and Cleaning

Nama Fitur/Kolom	Tipe Data	Penjelasan	Masalah	Penyelesaian
Agency	Object	Badan yang menjual produk asuransi kepada nasabah (Frontliner)	-	-
Agency Type	Object	Tipe proteksi perjalanan	-	-
Distribution Channel	Object	Metode yang digunakan untuk berjualan	-	-
Product Name	Object	Paket asuransi yang memiliki kemanfaatan tersendiri	-	-
Gender	Object	Jenis kelamin nasabah	Terdapat missing values (69%)	Drop null values (karena dianggap anomali)
Duration	int	Lama nya waktu perjalanan	Outliers yang signifikan	Drop nilai yang < 180 hari
Destination	Object	Destinasi tujuan	-	-
Net sales	int	Jumlah netto penjualan produk oleh agen asuransi	Outliers yang signifikan	Tidak dilakukan treatment
Comision value	int	Komisi yang didapatkan agen asuransi saat berhasil menjual produk	Outliers yang signifikan	Tidak dilakukan treatment
Age	Int	Umur nasabah yang diasuransikan	Outliers yang signifikan	Drop values > 75 tahun
Claim	Object	Status klaim	Value kolom berupa string (Yes dan No)	Mapping Nilai 1 dan 0

Total Data Before Cleaning

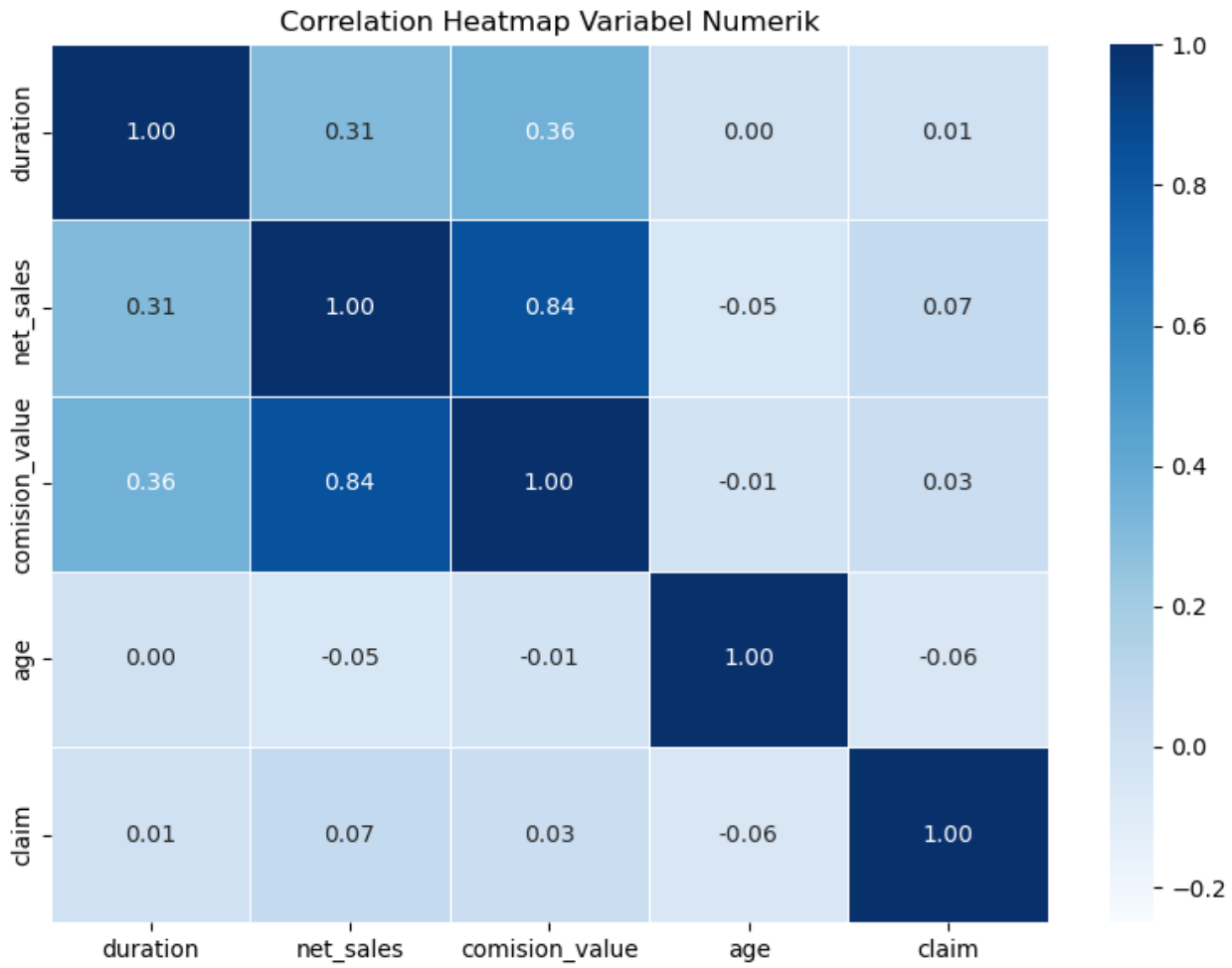
44,328

Total Data After Cleaning

10,039

Data Understanding and Cleaning

Korelasi antar variable numerik



Top Variable Kategorik

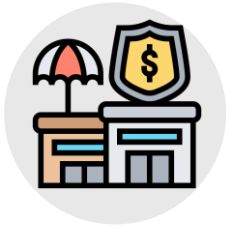
Agency

C2B

Agency Type

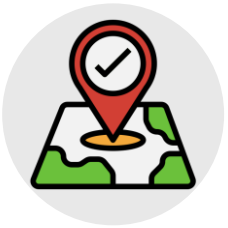
Airlines

Product Name



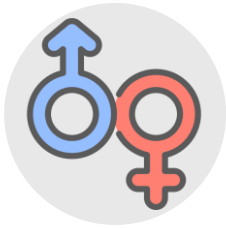
Basic Plan

Destination



Singapore

Gender



Male

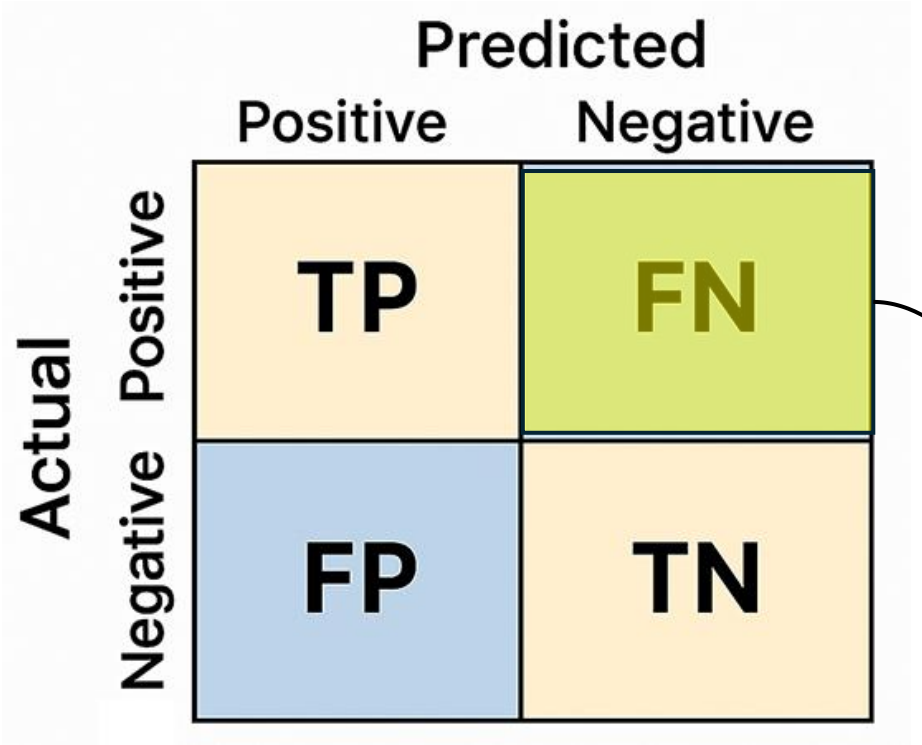
Claim

No

Preprocessing

01	**Scaling**	[‘duration’, ‘net_sale’,’comision_value]
02	**Encoding**	**One Hoot Encoding** [‘gender’, ‘distribution_chanel’,agency_type] **Binary Encoding** [‘agency’, ‘product_name’,destination]
03	**Resampling**	ROS (Random Oversampling) SMOTE Penalized Method

Evaluation Method



resiko **salah melakukan prediksi nasabah/pemegang polis** yang tidak melakukan klaim namun secara actual melakukan klaim memiliki dampak resiko tinggi

Model Benchmarking

Model	AVG Recall Score	STD Recall
LogisticRegression (ROS)	0.876471	0.046407
LogisticRegression (smote)	0.864932	0.043663
SVM (ROS)	0.799321	0.039158
SVM (smote)	0.760558	0.038034
Gradient Boosting (ROS)	0.640649	0.078955
Gradient Boosting (smote)	0.478356	0.083644
KNN (smote)	0.339744	0.049321
LG Boosting (ROS)	0.262594	0.079514
DecisionTreeClassifier (smote)	0.146757	0.026291
KNN (ROS)	0.139065	0.044747
RandomForestClassifier (smote)	0.131373	0.046470
XG Boosting (ROS)	0.123605	0.023363
DecisionTreeClassifier (ROS)	0.088688	0.025822
XG Boosting (smote)	0.081146	0.028478
LG Boosting (smote)	0.034842	0.019123
RandomForestClassifier (ROS)	0.026998	0.009361

Modeling Dataset

Model	Recall Train	Recall Test	Delta Reesult (Train - Recall)
Logistic Regression (Smote)	0.864932	0.876923	-0.011991
Logistic Regression (Panalized)	0.876471	0.892308	-0.015837
SVM Panalized	0.810860	0.830769	-0.019910

Hyperparameter Tuning

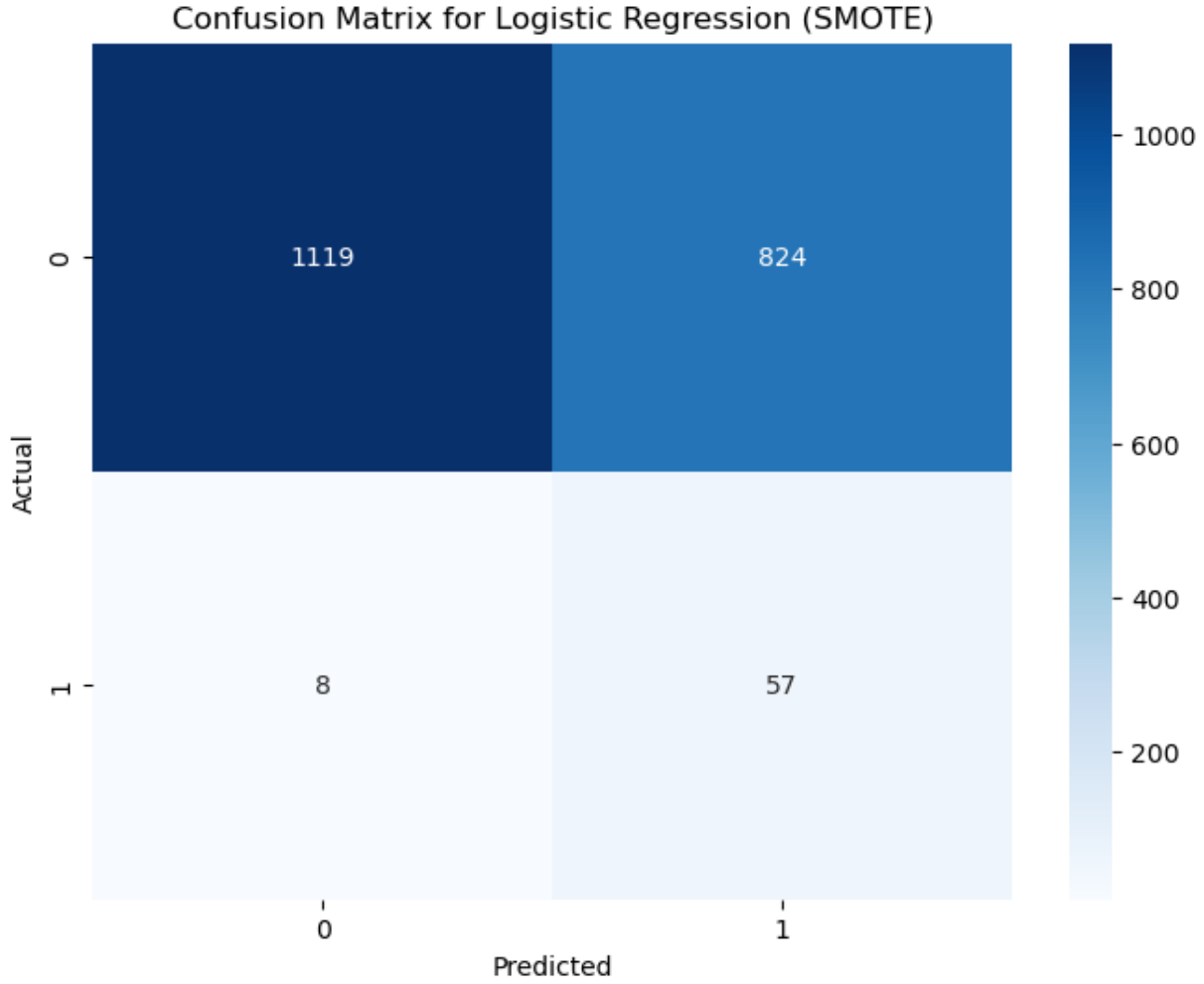
Model	Recall Train	Recall Test	Recall Train (Tuned)	Recall Test (Tuned)	Delta Result (Train - Recall) tuned
Logistic Regression (Smote)	0.864932	0.876923	0.864932	0.876923	-0.011991
Logistic Regression (Panalized)	0.876471	0.892308	0.876471	0.892308	-0.015837
SVM Panalized	0.810860	0.830769	0.818703	0.846154	-0.027451

Evaluation

Feature Importance

Berdasarkan feature importance logistic regression yang dilihat dari koefisien beta di setiap fitur, terlihat bahwa **distribusi_channel, destinasi, dan agency_type** memberikan pengaruh terhadap probabilitas data

Confusion Matrix



Kemampuan model dalam menangkap kategori 1 (klaim) menghasilkan skor 88% namun terlihat bahwa kemampuan prediksi kemungkinan benar ada model hanya sebesar 6%

Conclusion

1. Model terbaik yang digunakan dalam kasus ini Adalah Logistic Regression SMOTE
2. Peformance Recall Adalah 87%
3. Distribution_channel, destinasi, dan agency_type merupakan fitur yang penting
4. Berdasarkan cost benefit analysis, total cost yang dikeluarkan menggunakan Machine learning sebesar $242,000

Recommendation

Machine Learning Model

1. Lakukan training ulang secara berkala untuk menyesuakian pola data pemegang polis

2. Monitor metriks penilaian secara berkala untuk menjaga konsistensi prediksi

3. Mengaplikasikan metode lain untuk imbalance data dan feature engineering

Business

1. **Penentuan premi berdasarkan faktor resiko** - Model dapat digunakan untuk menentukan kategoriasi resiko pemegang polis:
- Prediksi klaim tinggi -> sesuaikan premi untuk mengcover cost.
- Prediksi klaim tinggi -> sesuaikan premi dengan harga yang cukup rendah dan kompetitif di pasar.

2. **Segmentasi pasar**
- Segmentasi dapat dilakukan melihat dari tipe resikonya.
- Fokus kepada **low-medium risk** dikarenakan segmen mayoritas pemegang polis tidak melakukan klaim

3. **Operational Risk Management**
-Alokasi dana assessment untuk keperluan penetrasi kepada calon pemegang polis yang beresiko tinggi.