Purwadhika
Digital Technology School

Telco Customer Churn Prediction

Oleh :
Muhamad Kahfi Dwi Prasetio

Outline

- Pendahuluan
- Analisis Data
- Machine Learning Modelling
- Simulasi Prediksi Data Churn Untuk Manajer Pemasaran
- Kesimpulan dan Rekomendasi

Pendahuluan

Latar Belakang

Manajemen Pemasaran sebuah perusahaan telekomunikasi menghadapi tantangan besar terkait tingkat churn pelanggan yang cukup tinggi dan mengalami kerugian sebesar $71,300. Mereka selama ini mengeluarkan biaya promosi ke seluruh pelanggan tanpa tahu siapa yang benar-benar berisiko berhenti menggunakan layanan. Strategi ini tidak efisien, karena sebagian besar promosi justru jatuh kepada pelanggan yang sebenarnya tidak berniat berhenti berlangganan.

Rumusan Masalah

Perusahaan belum memiliki sistem yang mampu memprediksi kemungkinan pelanggan akan berhenti berlangganan. Akibatnya, strategi promosi dilakukan secara massal ke seluruh pelanggan, yang mengakibatkan biaya promosi yang besar namun tidak efisien.

Metric Evaluation

Cost FP: $100

Cost FN: $500

- **False Positive (FP)**: Model memprediksi akan churn → beri promosi $100 Tapi kenyataannya pelanggan tidak akan churn → promosi tidak perlu → rugi $100
- **False Negative (FN)**: Model memprediksi tidak akan churn → tidak beri promosi Tapi kenyataannya pelanggan churn → kehilangan pelanggan → rugi $500

Karena cost dari FN jauh lebih tinggi dibanding FP, karena itu **F2-score** dipilih sebagai metrik evaluasi utama.

Tujuan Analisis

- Memprediksi Churn Pelanggan
- Mengurangi Kerugian Finansial
- Mengoptimalkan Strategi Promosi
- Menggunkan Metode Evaluasi yang Tepat
- Memberikan Insight Fitur Penting

Kondisi Data

- Data awal pada file **data_telco_customer_churn.csv** mempunyai **4930 rows**.
- Data memiliki 11 kolom yang terdiri dari 10 kolom feature dan 1 kolom target, kolom **target** pada data ini adalah **Churn.**
- Tipe datanya **9 bertipe Kategorikal** dan **2 bertipe Numerikal**.
- Perbandingan pelanggan yang Churn dan yang tidak Churn adalah **1288 : 3565**

Kolom Feature

- Dependent
- Tenure
- OnlineSecurity
- OnlineBackup
- InternetService
- DeviceProtection
- TechSupport
- Contract
- PaperlessBilling
- MonthlyChanges

Analisis Data

Analisis Data

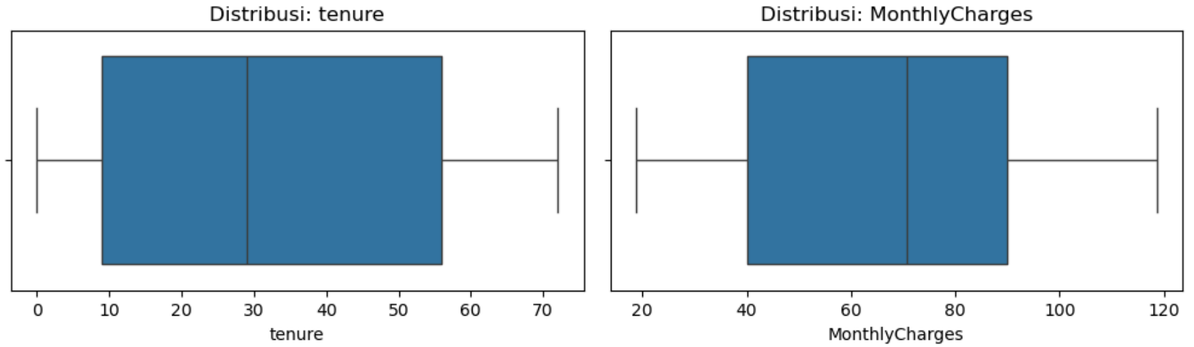
- Perbedaan pelanggan yang Churn hampir **4 kali** dengan pelanggan yang tidak Churn
- Perbandingan persentase pelanggan yang Churn dan yang tidak Churn adalah **27% : 73%**
- Dilihat dari **Pvalue** semua feature mempunyai kaitan dengan target (Churn)

Data Cleaning

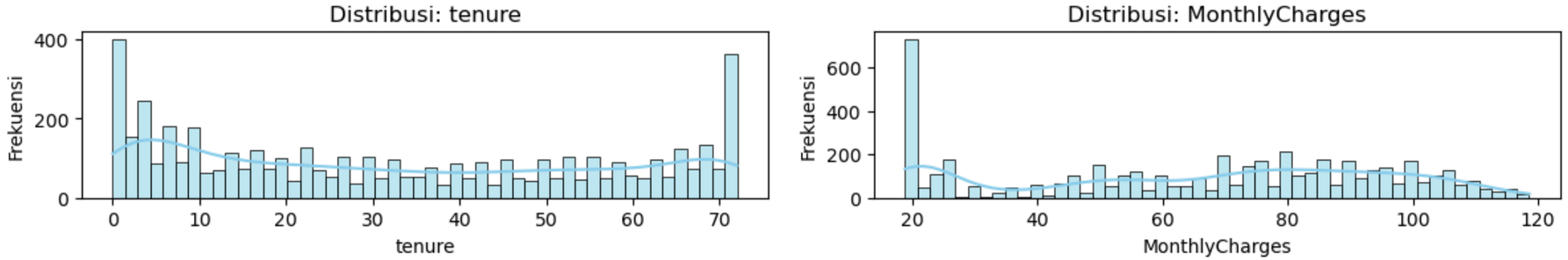
- Setelah melakukan **pengecekan missing values** ternyata tidak ada value yang hilang atau berisi Nan,
- **Pengecekan outlier**, tidak ada outlier dari kolom numerikal (tenure dan MonthlyCharges)
- Tidak ada value **kolom Kategorikal** yang dihapus atau digabungkan karena value countnnya sedikit
- **Pengecekan duplikat**, ternyata data memiliki dupikat sebanyak **77** data dan didrop jumlah row akhir data menjadi **4853**.
- Merubah value kolom target (**Churn**) dari **Yes → 1 dan No → 0**

Data Numerikal

Cek Outlier Feature Numerical



Cek Normalitas Feature Numerical



Machine Learning Modeling

Tahapan Machine Learning

- Define X and Y
- Data Splitting
- Data Preprocessing
- Hyperparameter Tuning
- Performance in Test Set
- Best Model
- Confusion Matrix
- Feature Importance

Define X and y

- **Feature**:'Dependents', 'tenure', 'OnlineSecurity', 'OnlineBackup', 'InternetService', 'DeviceProtection', 'TechSupport', 'Contract', 'PaperlessBilling', dan 'MonthlyCharges'

- **Target:** 'Churn'

```
X = df.drop(['Churn'], axis=1)
y = df['Churn']
```

Data Splitting

```
X_train, X_test, y_train, y_test = train_test_split(
  X,
  y,
  stratify = y,
  test_size = 0.2,
  random_state = 0 )
```

- stratify: Menjaga proporsi distribusi kelas antara training dan testing set agar tetap sama seperti data asli
- test_size: Menentukan berapa besar proporsi data yang akan digunakan sebagai test set.
- random_state: Menetapkan seed angka acak agar pembagian data selalu konsisten setiap dijalankan (reproducibility).

Data Preprocessing

- **Encoding** : **OneHot** : 'Dependents', 'OnlineSecurity' 'OnlineBackup', 'InternetService' 'DeviceProtection', 'TechSupport', 'Contract', 'PaperlessBilling'
- **Scaling** : **Robust** : 'tenure', 'MonthlyCharges'

Top 3 Model Terbaik

MODEL	MEAN	STD
GradienBosst	**0.709358**	**0.012377**
LogReg	**0.705890**	**0.015714**
XGBoost	**0.612422**	**0.009265**

Hyperparameter Tuning

Logreg

- Best_score: **0.7239371836742807**
- Best_params: {*'resampler': RandomOverSampler(random_state=42), 'model__solver': 'liblinear', 'model__penalty': 'l1', 'model__C': 0.01*}

GradienBoost

- Best_score: **0.7298450434156383**
- Best_params: {*'resampler': RandomUnderSampler(random_state=42), 'model__subsample': 0.3, 'model__n_estimators': 234, 'model__max_features': 5, 'model__max_depth': 1, 'model__learning_rate': 0.2*}

Performance in Test Set

Before Hyperparameter Tuning

- Running model: LogisticRegression F2 Score: **0.7203**
- Running model: GradientBoostingClassifier F2 Score: **0.7066**

After Hyperparameter Tuning

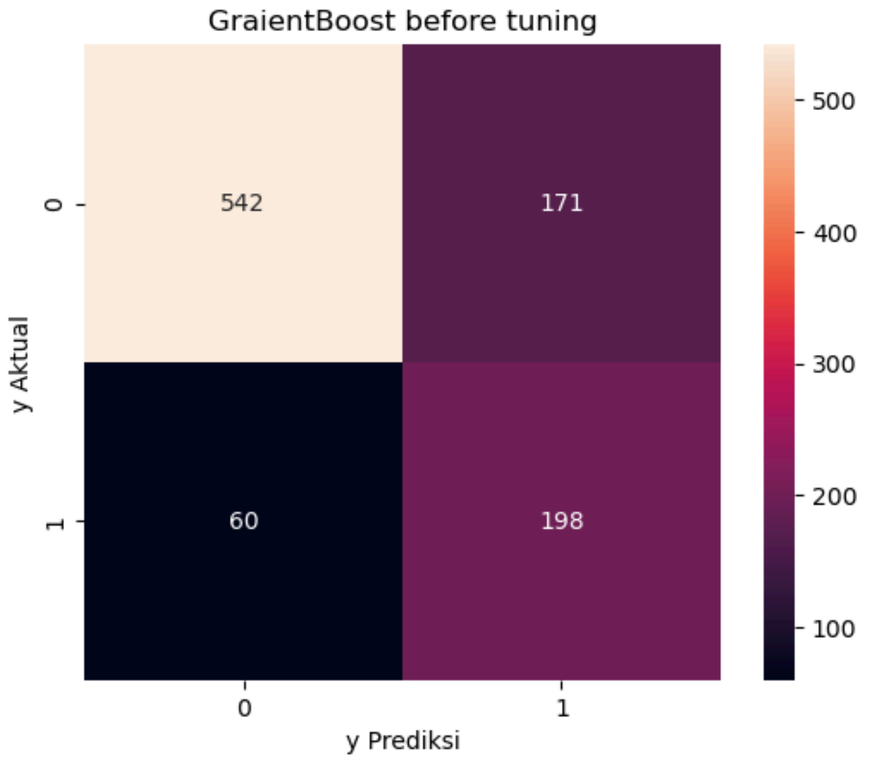
- Evaluating: LogisticRegression (Tuned) F2 Score: **0.7182**
- Evaluating: GradientBoostingClassifier (Tuned) F2 Score: **0.7554**

Best Model

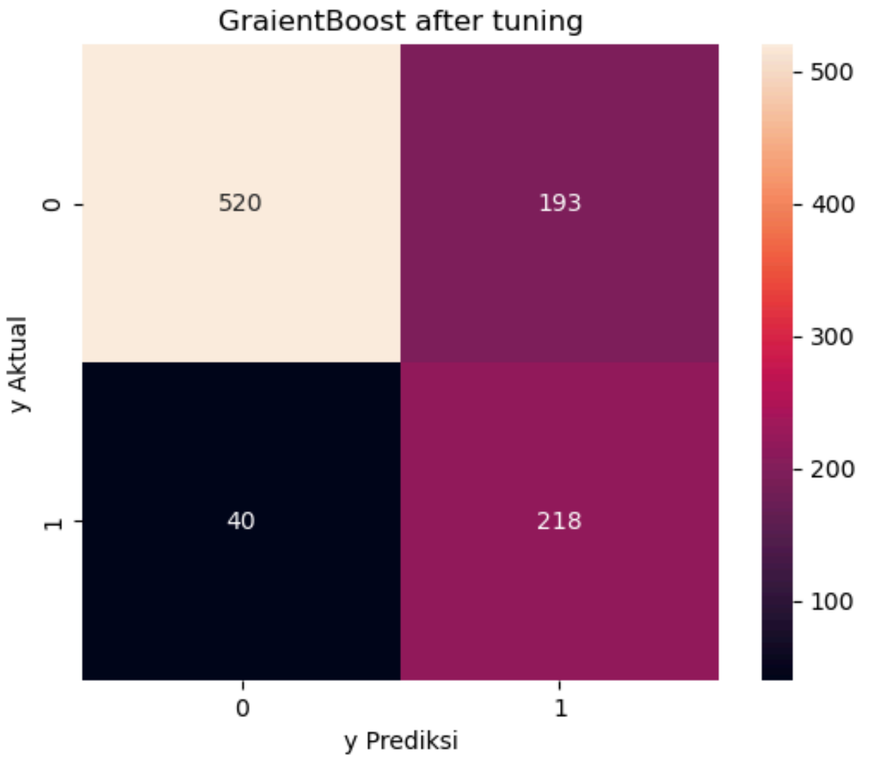
Gradient Boosting setelah tuning adalah pilihan yang lebih optimal:

- **Best_params**:*{'resampler': RandomUnderSampler (random_state=42), 'model__subsample': 0.3, 'model__n_estimators': 234, 'model__max_features': 5, 'model__max_depth': 1, 'model__learning_rate': 0.2}*
- **model_pipe =** *Pipeline ([('transformer', transformer), ('balancing', SMOTE (random_state=0), ('model', GradientBoostingClassifier (random_state=0)])*
- F2 Score (**Train Set**): 0.729
- **Before tuning** (**Test Set**) GradientBoostingClassifier : 0.71
- **After tuning** (**Test Set**) GradientBoostingClassifier : 0.755

Confusion Matrix



- Model cukup baik mengenali pelanggan churn, namun masih terdapat jumlah FN yang signifikan (60).
- Jumlah FP juga tinggi (171),



- Setelah tuning, model lebih baik dalam mendeteksi pelanggan yang benar-benar churn (TP naik, FN turun).
- Namun, ada kenaikan FP

Prediksi Menggunakan ML vs Tanpa ML

Tanpa menggunakan Machine Learning

	Predicted (0)	Predicted (1)
Actual(0)	0	713
Actual(1)	0	258

Total biaya promosi (seluruh pelanggan):

- 971 × $100 = $97,100

Promosi yang tepat sasaran (untuk 258 pelanggan churn):

- 258 × $100 = $25,800

Biaya promosi yang sia-sia ke pelanggan loyal:

- 713 × $100 = **$71,300**

Dengan menggunakan Machine Learning

	Predicted (0)	Predicted (1)
Actual(0)	520	193
Actual(1)	40	218

False Positive (FP):

- 193 × $100 = $19,300 → biaya promosi ke pelanggan loyal

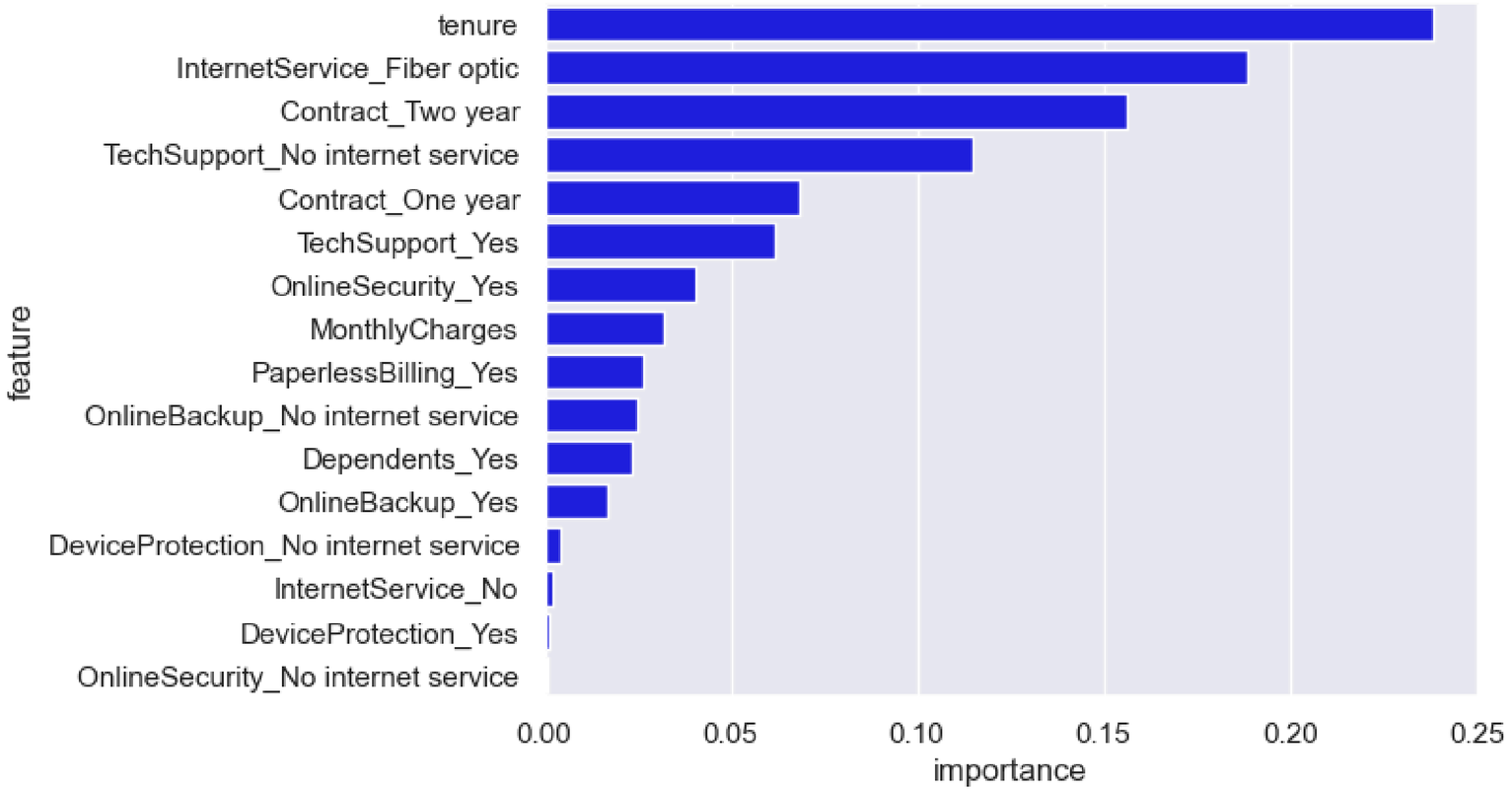
False Negative (FN):

- 40 × $500 = $20,000 → kehilangan customer karena tidak dipromosikan

Total kerugian:

- $19,300 + $20,000 = **$39,300**

Feature Importance



Berdasarkan feature importance dari model Gradient Boosting, fitur yang paling berpengaruh terhadap churn adalah:

- Tenure
- Contract
- InternetService

Simulasi Prediksi Data Churn Untuk Manajer Pemasaran

SCAN ME

Kesimpulan dan Rekomendasi

Kesimpulan

Tingkat Churn Pelanggan masih tergolong tinggi yakni sebesar 27% yang merupakan sinyal managemen untuk meningkatkan retensi pelanggan

Model Gradient Boosting

- F2 Score (Train Set): 0.729
- F2 Score (Test Set): 0.755

Pengaruh Feature terhadap Churn dimana feature yang paling berpengaruh adalah tenure, Contract, dan MontlyChages

Machine Learning berhasil menurunkan kerugian sampai 45% dari total kerugian tanpa Machine Learning

Rekomendasi

1. Implementasikan Model Gradient Boosting ke Proses Operasional

2. Fokus pada Fitur Penting dan Pertimbangkan Penghapusan Fitur Kurang Relevan

3. Kembangkan Strategi Customer Retention

4. Monitoring & Evaluasi Berkala

Sekian dan Terima Kasih