

Final Project Presentation

Nomor Kelompok: 3

Nama Mentor: Aditya Bariq

Nama:

- Muhammad Rizqiansyah
- Nadia Rizky Hairunnisa

Machine Learning Class

Program Studi Independen Bersertifikat
Zenius Bersama Kampus Merdeka



- 1. Latar Belakang**
- 2. Eksplorasi Data dan Visualisasi**
- 3. Modelling**
- 4. Kesimpulan**



Latar Belakang

Latar Belakang Project

Sumber Data: Walmart Dataset

<https://www.kaggle.com/datasets/yasserh/walmart-dataset>

Problem: **Regression**

Tujuan:

- Memprediksi penjualan mingguan di Walmart

Eksplorasi Data dan Visualisasi



Business Understanding

Walmart merupakan salah satu perusahaan *retail* multinasional terbesar di dunia. Walmart memiliki banyak pesaing yang bergerak di bidang *retail* sehingga diperlukan keputusan yang strategis agar bisa mempertahankan posisinya.



Business Understanding

Resource/Dataset:

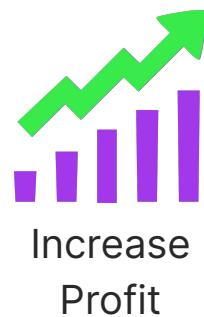
- Gabungan data dari 45 toko termasuk informasi toko dan penjualan mingguan.
- Data disediakan setiap minggu
- Terdapat 4 minggu liburan (Natal, Thanksgiving, Super bowl, Hari Buruh)



Business Understanding

Business Objectives:

- Apakah terdapat insights pada data? Sehingga kita bisa...



Side Questions: Bagaimana faktor waktu dan perekonomian negara bisa mempengaruhi Weekly Sales?

Data Cleansing

Dimensi data: **6435 baris dan 8 kolom**

Kolom target: **Weekly_Sales**

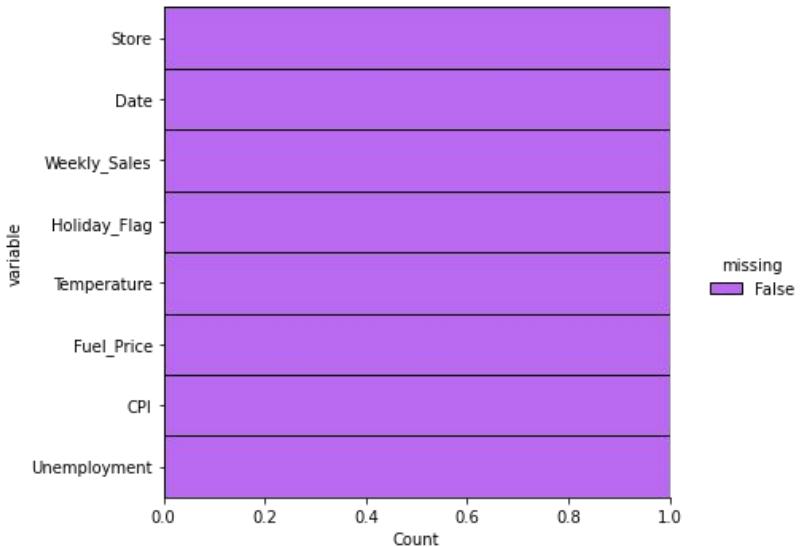
Missing values: **0**

Duplicated values: **0**

Jumlah toko: **45**

Hari unik: **Jum'at***

*Pencatatan data dilakukan setiap hari Jum'at



Data Cleansing

Info Hari Libur Besar dari tahun 2010-2012:

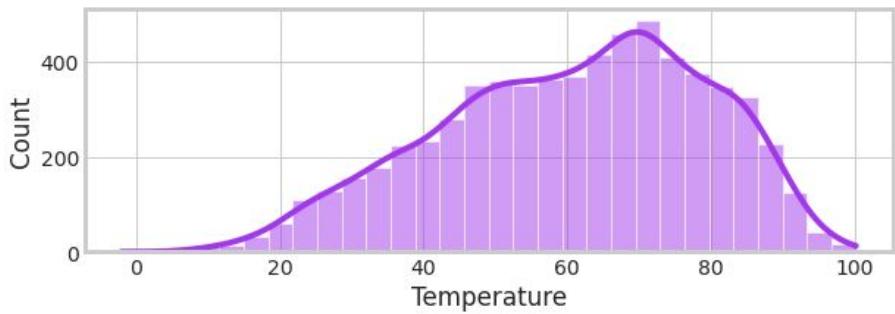
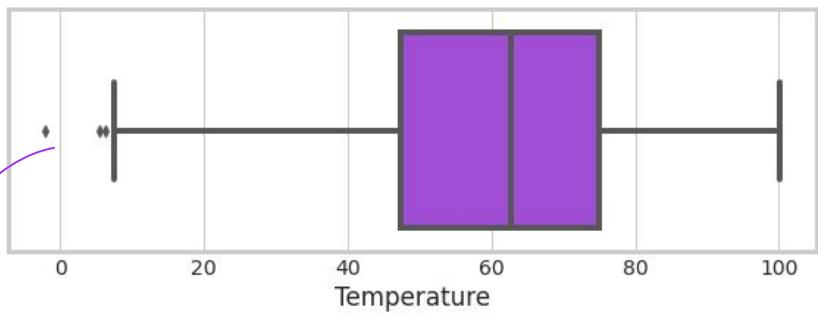
- Super Bowl: Seluruhnya terdapat dalam dataset
 - Labor Day: Seluruhnya terdapat dalam dataset
 - Thanksgiving: **Tidak ada data tahun 2012***
 - Christmas: **Tidak ada data tahun 2012***
- 
- The diagram consists of two wavy purple lines pointing from the asterisk (*) in the Thanksgiving and Christmas entries to explanatory text. The top line points to the text "Harusnya ada di bulan November 2012" (Should have been in November 2012). The bottom line points to the text "Harusnya ada di bulan Desember 2012" (Should have been in December 2012).

Range data: **5 Februari 2010 - 6 Oktober 2012**

*akibat dari range data yang kurang lengkap. Hari libur besar tersebut berada di luar range data

Data Cleansing

Outliers: Kolom Temperature = 3 data



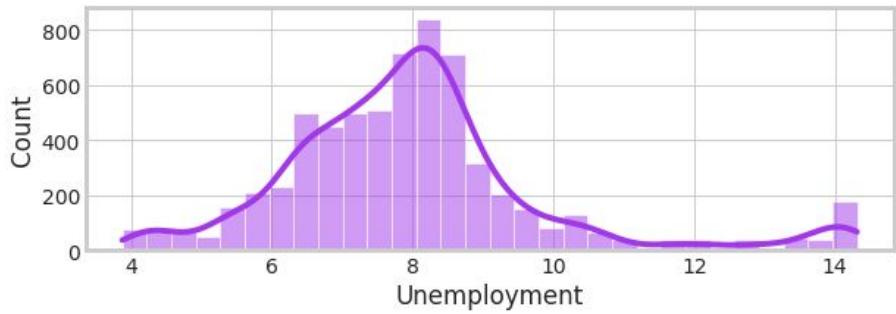
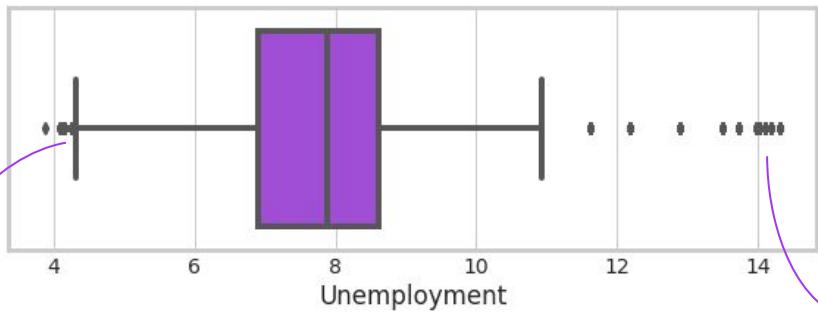
Wajar. Outliers ada di suhu terendah dan berada di Bulan Januari dan Februari.

Asumsi: Data temperatur pada hari itu diambil ketika musim dingin

Data Pendukung: Pada tahun 2011, musim dingin dimulai dari awal Desember 2010 dan berakhir di akhir Februari 2011

Data Cleansing

Outliers: Kolom Unemployment = **481** data

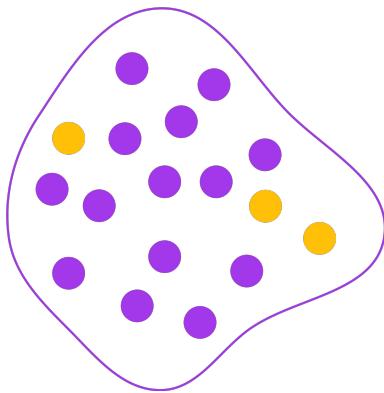


Unemployment **rendah** terdapat pada tahun **2012**. Sementara itu, Unemployment **tinggi** terdapat pada **semua tahun**. Sehingga, **perlu dilakukan handling outlier lebih jauh**

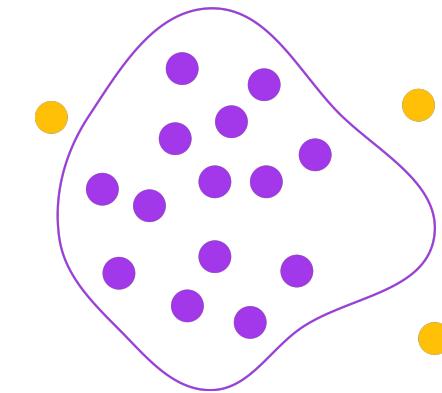
Handling Outliers

3 Skenario handling outliers

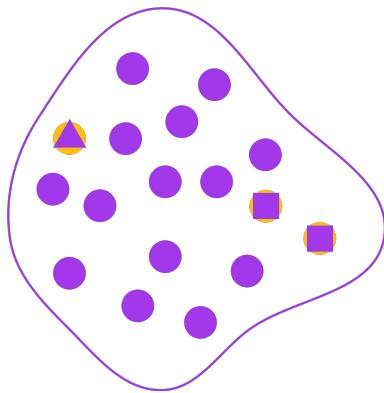
Outlier dibiarkan



Outlier dihapus

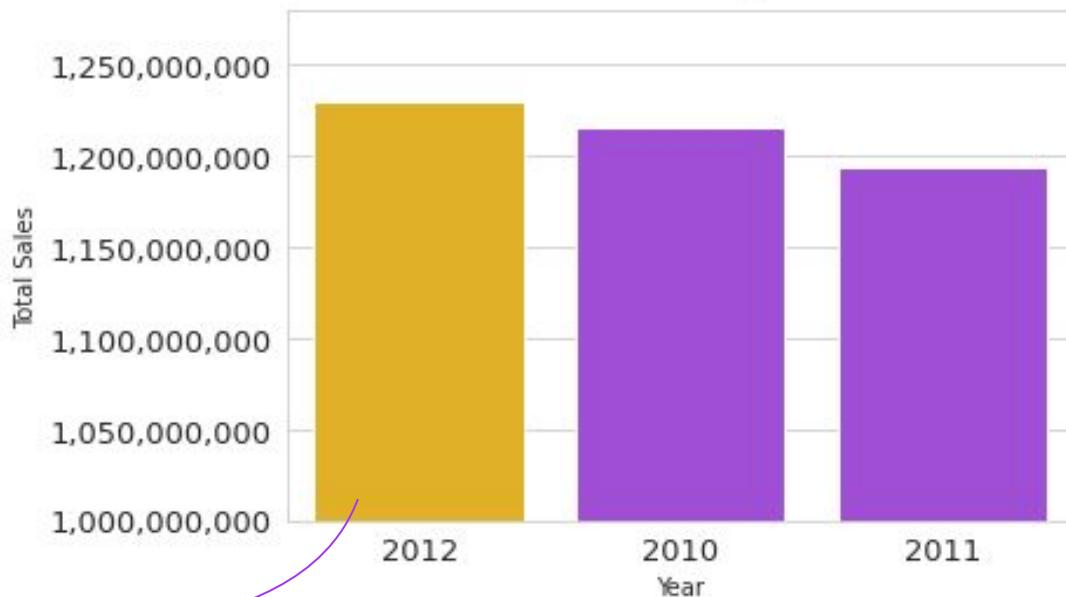


Transformasi outlier



Exploratory Data Analysis

Annual Sales during 2010-2012



Insight: Annual Sales terbesar terjadi pada tahun **2012**.

Notes: Karena range data setiap tahun tidak seimbang, Total Weekly Sales dihitung dari bulan **Februari hingga Juli** saja untuk melihat perbandingannya.

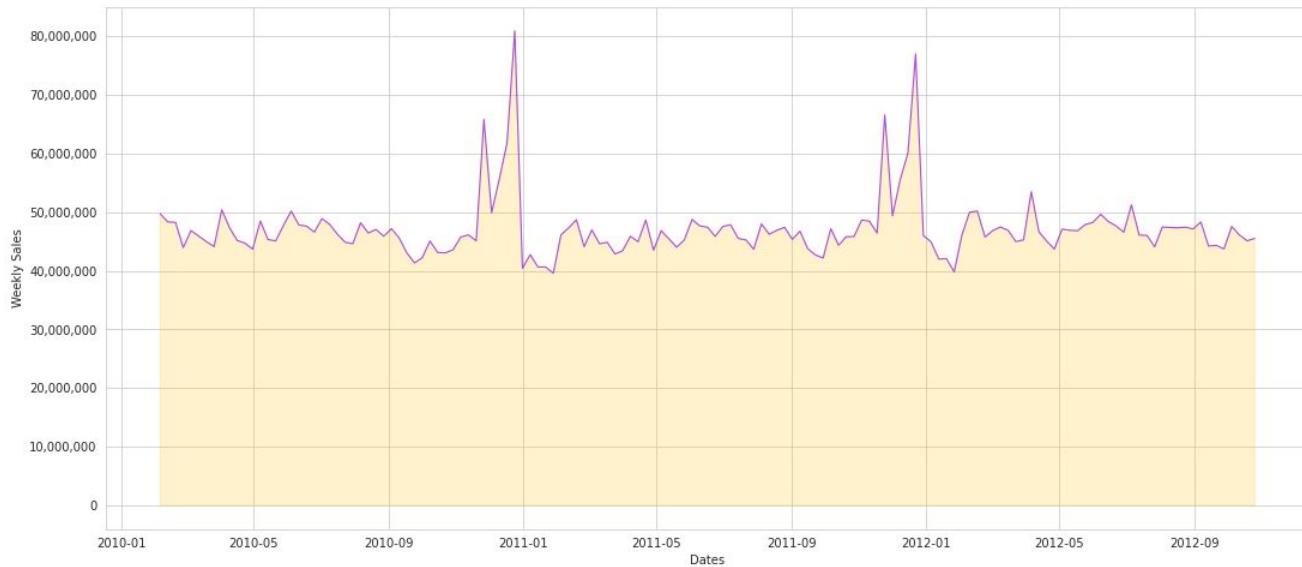
Exploratory Data Analysis

Insight: Average Weekly Sales terbesar terjadi pada bulan Desember.



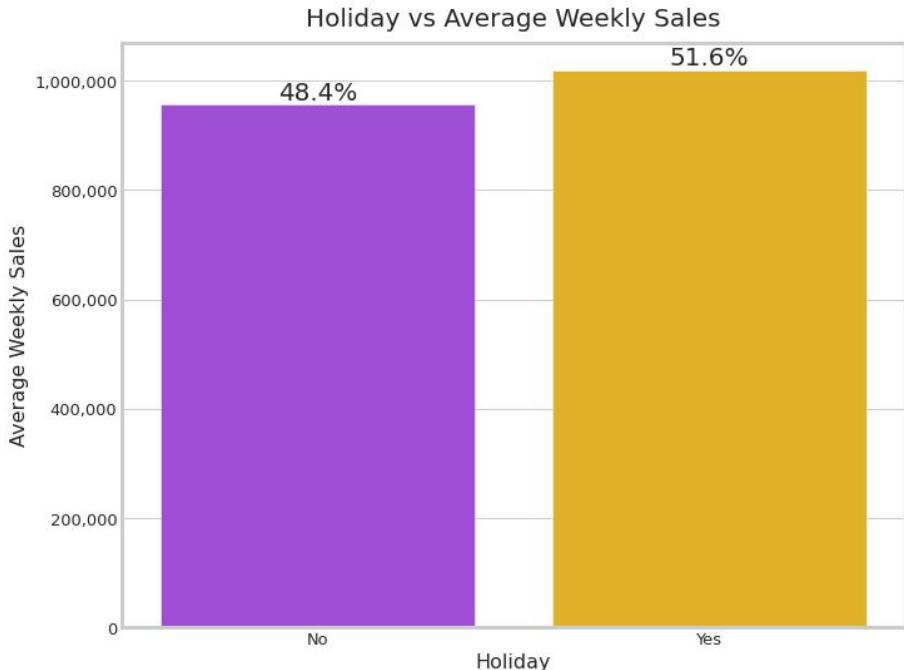
Exploratory Data Analysis

Weekly Sales Everyday



Insight: Terdapat pola tahunan dimana penjualan meningkat signifikan di bulan **November** dan **Desember**.

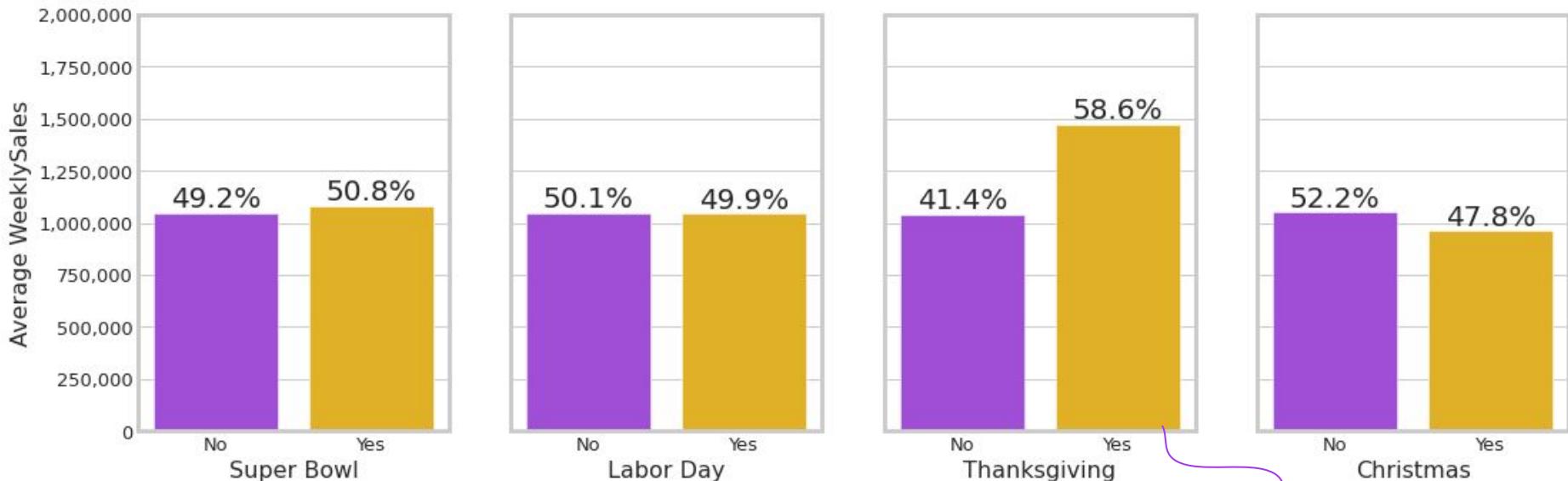
Exploratory Data Analysis



Insight: Terdapat selisih Average Weekly Sales sebesar **3,2 %** lebih tinggi pada **Holiday** dibandingkan **Hari Biasa**.

Exploratory Data Analysis

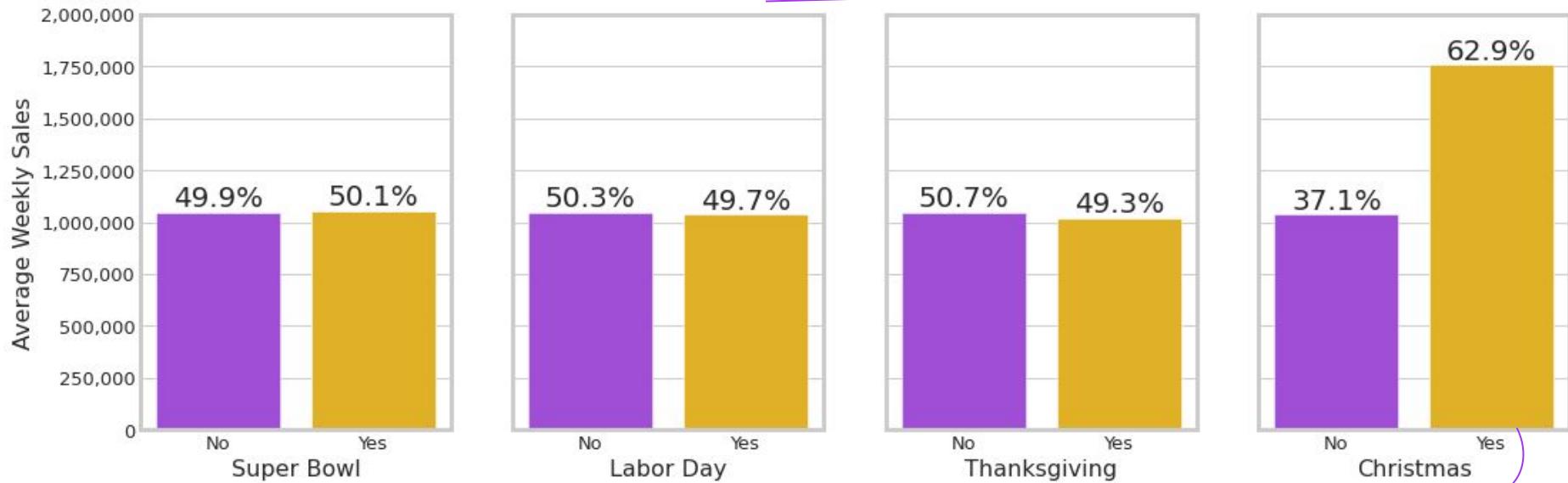
Average Weekly Sales on Big Holiday Week



Insight: Average Weekly Sales tertinggi ketika Holiday Week terjadi di **Thanksgiving**.
Dimana perbandingannya sebesar **17.2 %** lebih tinggi dibandingkan Hari Biasa

Exploratory Data Analysis

Average Weekly Sales A Week Before Big Holiday Week

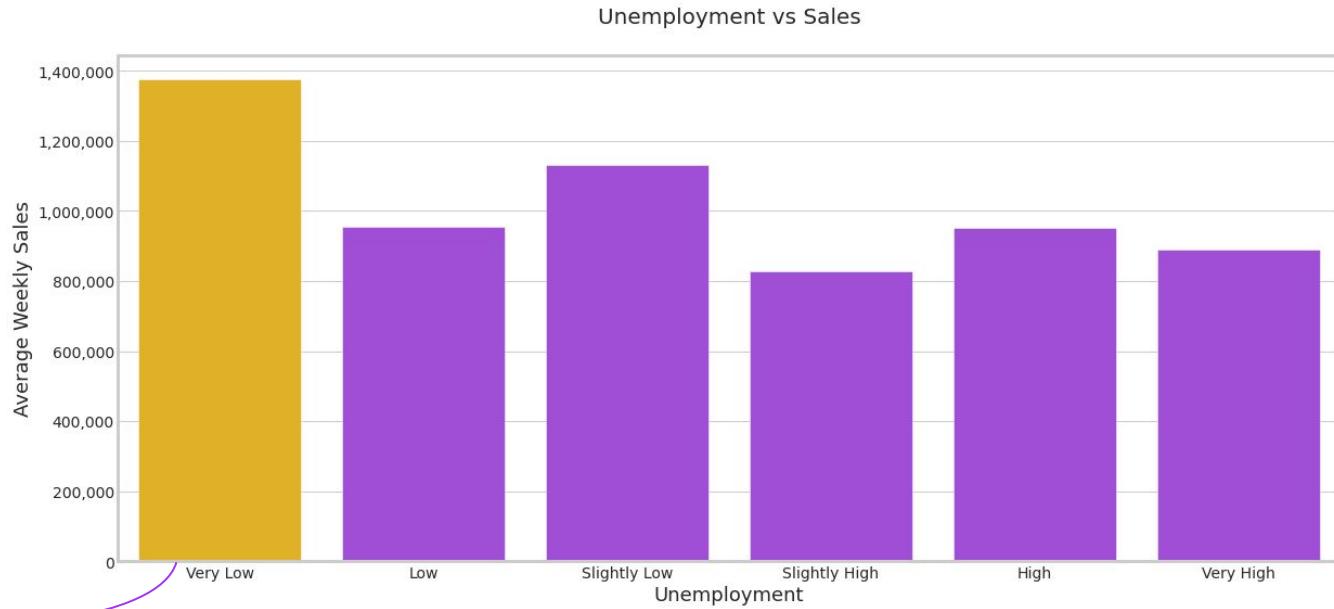


Insight: Average Weekly Sales tertinggi **1 minggu sebelum** Holiday Week terjadi di **Christmas**.
Dimana perbandingannya sebesar **25,8 %** dibandingkan Hari Biasa

Exploratory Data Analysis

Kategori Unemployment Rate:
8.0 - 9.9 : 1 (Very Low)
10.0 - 10.9 : 2 (Low)
11.0 - 11.9 : 3 (Slightly Low)
12.0 - 12.9 : 4 (Slightly High)
13.0 - 13.9 : 5 (High)
14.0 - 14.9 : 6 (Very High)

Insight: Average Weekly Sales tertinggi terjadi ketika Unemployment Rate Sangat Rendah.



Exploratory Data Analysis

Kategori CPI:

211.0 - 211.4 : 1 (Very-very Low)

211.5 - 211.9 : 1 (Very Low)

212.0 - 212.4 : 2 (Low)

212.5 - 212.9 : 3 (Slightly Low)

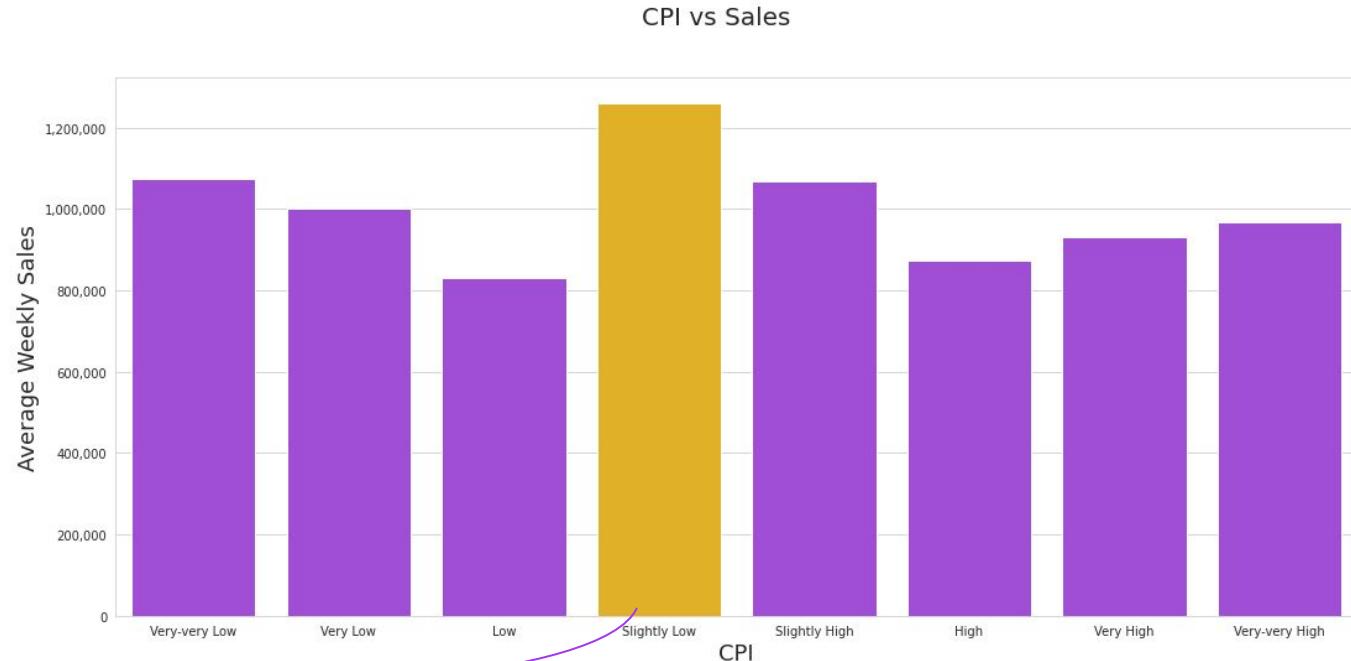
213.0 - 213.4 : 4 (Slightly High)

213.5 - 213.9 : 5 (High)

214.0 - 214.4 : 6 (Very High)

214.5 - 214.9 : 6 (Very High)

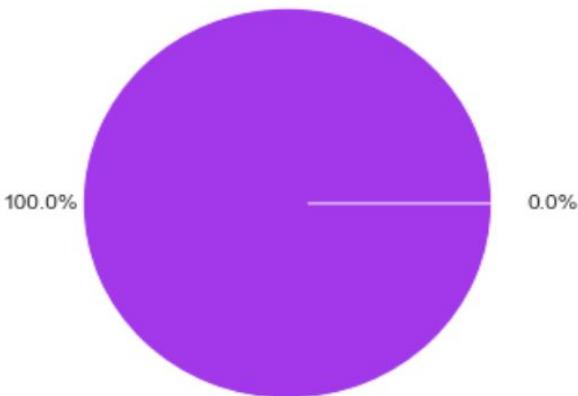
Insight: Average Weekly Sales tertinggi terjadi ketika CPI Sedikit Rendah.



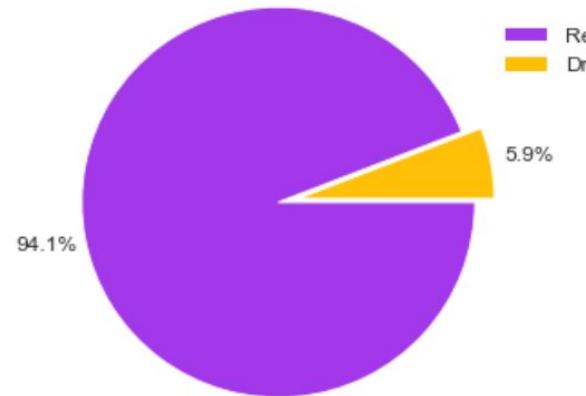
Modelling

Final Dataset

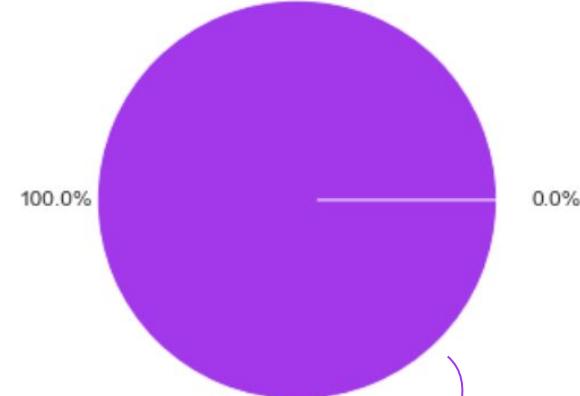
Dataset 1 (Without Dropping Outliers)



Dataset 2 (Outliers Dropped)



Dataset 3 (Outliers Flooring & Capping)



Floor: 10th percentile
Cap: 90th percentile

Feature Selection & Decomposition

1. Recursive Feature Elimination (RFE)

menyeleksi fitur berdasarkan estimator secara rekursif

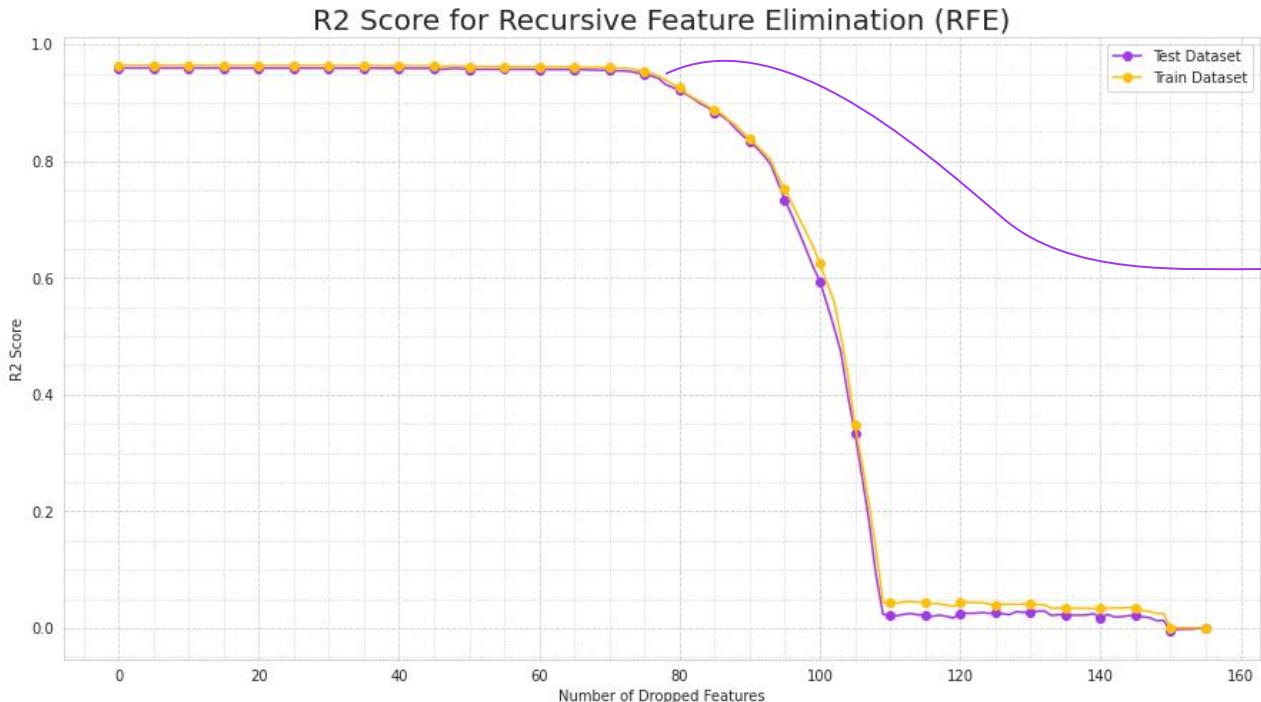
- (+) Mudah diinterpretasikan
- (-) Sangat bergantung pada estimator
- (-) Fitur yang berpotensi bisa saja tidak lolos seleksi

2. Principal Component Analysis (PCA)

mereduksi fitur dengan melakukan kombinasi berdasarkan nilai varians

- (+) Fitur akan dikombinasikan, bukan dieliminasi
- (-) Hasil sulit diinterpretasikan

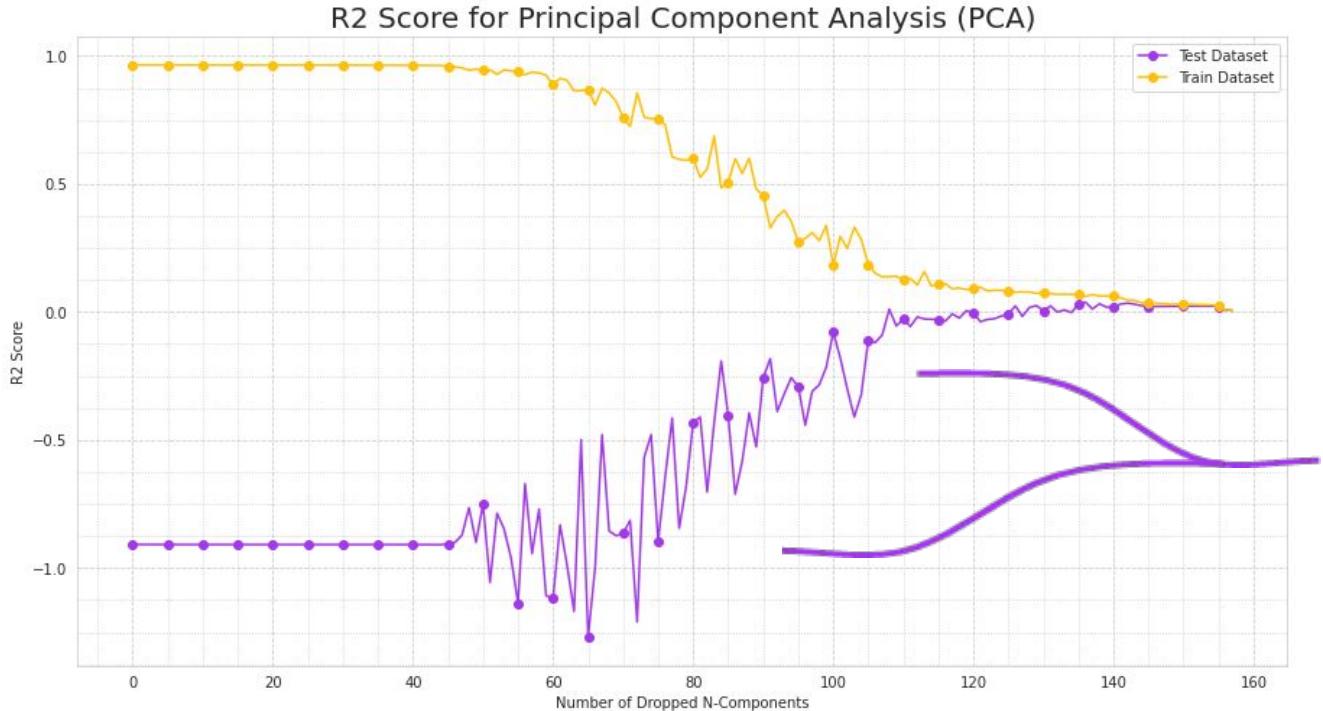
Feature Selection & Decomposition



Jumlah fitur optimal: 76

R2 Score mulai jatuh secara drastis setelah jumlah fitur yang di-drop melebihi 75, sehingga jumlah fitur yang terbaik adalah 76 (151-75)

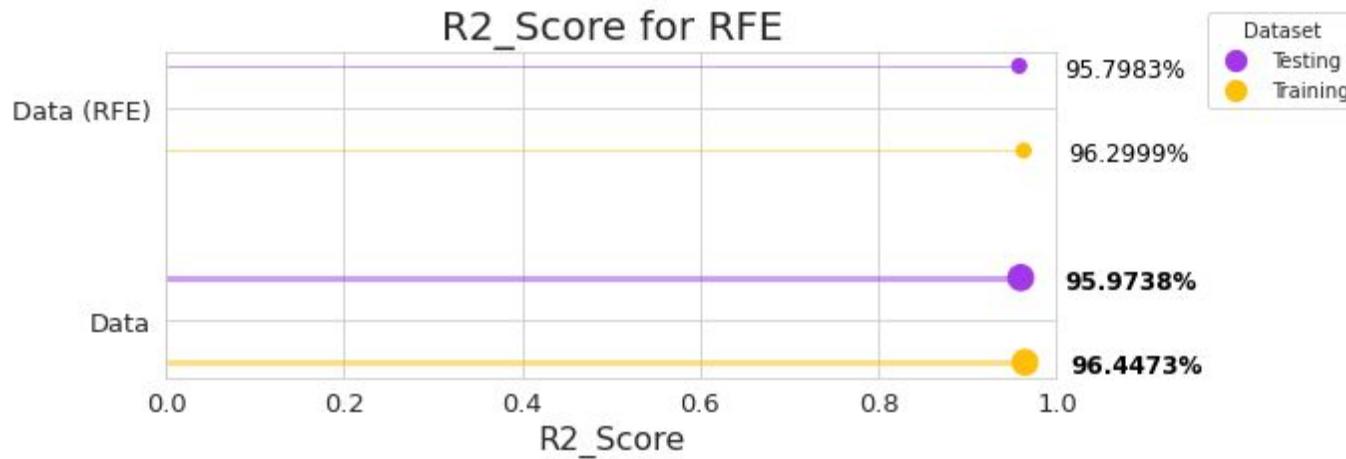
Feature Selection & Decomposition



PCA mengalami overfitting

Nilai R2 sangat baik pada training dataset, tetapi bernilai negatif dan naik turun secara drastis pada testing dataset

Feature Selection & Decomposition



Insight: Data tanpa seleksi fitur memiliki R2 score yang sedikit lebih baik

Setelah diukur waktu komputasinya, untuk 1x iterasi, proses training data tanpa RFE memakan waktu 0.128 s, sedangkan dengan RFE 0.124 s, selisih waktu 0.003s tidak terlalu signifikan karena jumlah datanya masih sedikit.

Model Selection

Dataset Splitting



```
graph LR; A[Dataset Splitting] --> B[Train: 60%]; A --> C[Test: 40%]
```

Model:

1. Linear Regression

Tidak ada penalti yang diberikan untuk tiap nilai bobot (weight).

2. Lasso Regression

Penalti diberikan sebesar jumlah dari nilai absolut bobot.

3. Ridge Regression

Penalti diberikan sebesar jumlah dari nilai kuadrat bobot.

4. ElasticNet Regression

Penalti diberikan menggunakan kombinasi Lasso dan Ridge.

Evaluation Metric

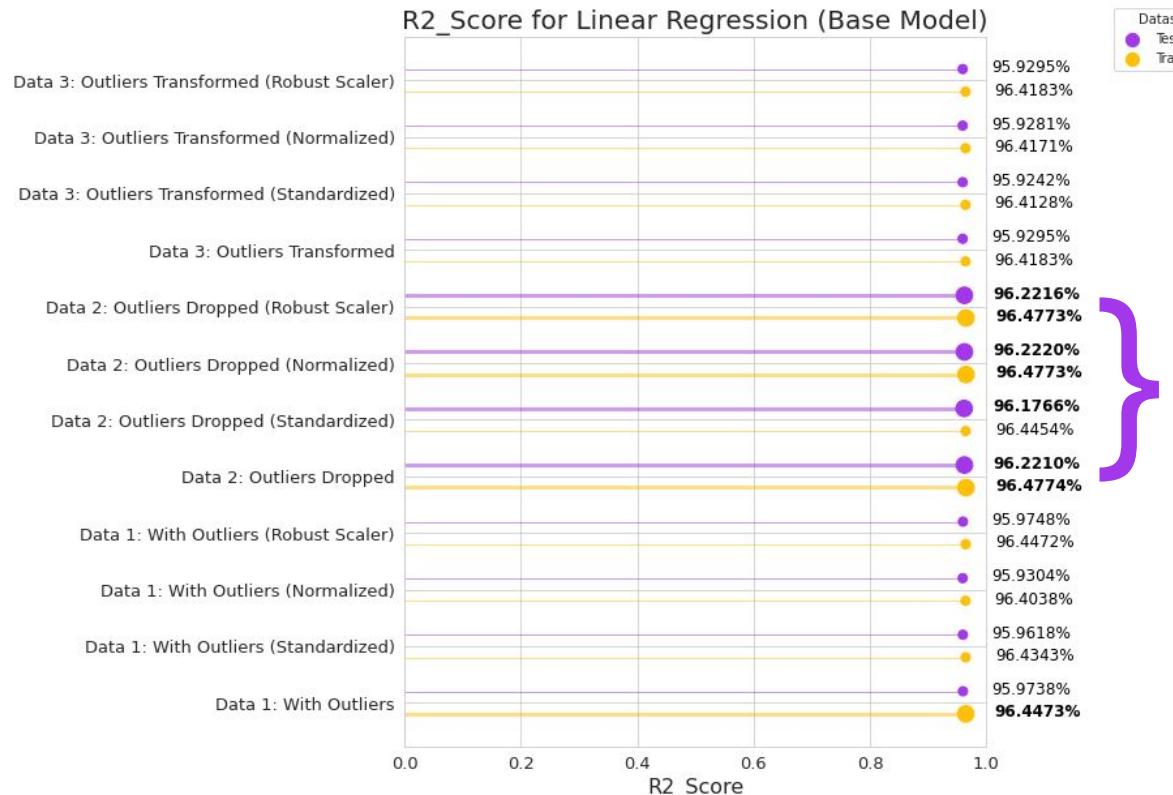
- R2-Score
 - MAE (Mean Absolute Error)
 - MSE (Mean Squared Error)
 - RMSE (Root Mean Squared Error)
 - MAPE (Mean Absolute Percentage Error)
 - **NRMSE (Normalized Root Mean Squared Error)**
-
- The diagram shows a vertical purple brace on the right side of the list, spanning from the RMSE entry up to the NRMSE entry. To the right of the brace, there are two rows of text. The top row contains '(+) widely used' and '(-) scale-dependent*'. The bottom row contains '(+) scale-independent' and '(-) asymmetric**'. A curved purple arrow points from the text 'RMSE/standard deviation' at the bottom left towards the brace.
- (+) widely used
(-) scale-dependent*
- (+) scale-independent
(-) asymmetric**

RMSE/standard deviation

* Supporting detail: [End-to-End Introduction to Evaluating Regression Models - AnalyticsVidhya](#)

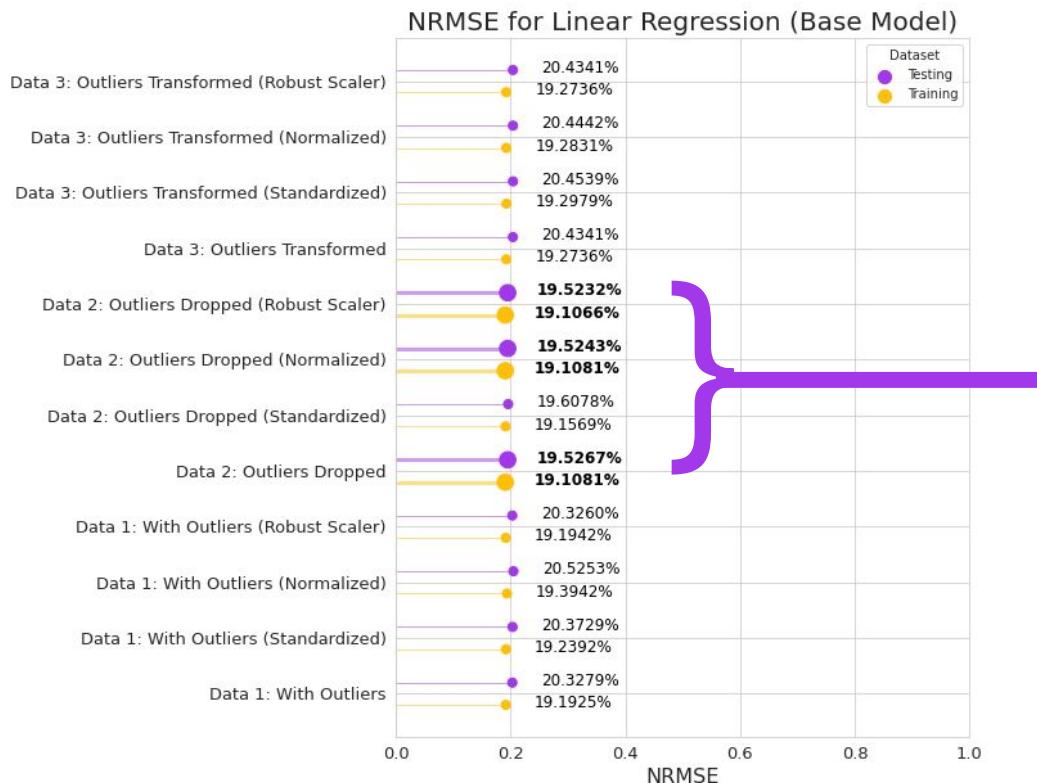
** Stack Exchange: [Is MAPE a good error measurement statistics?](#)

Model: Linear Regression



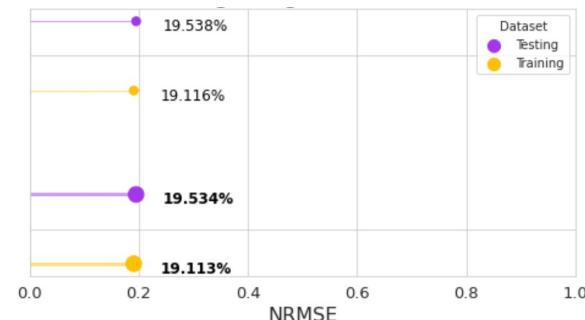
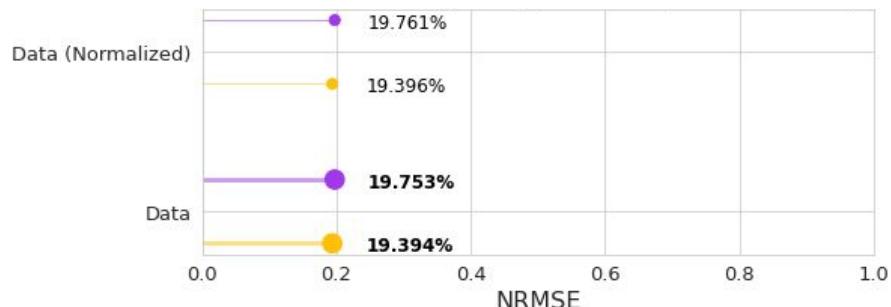
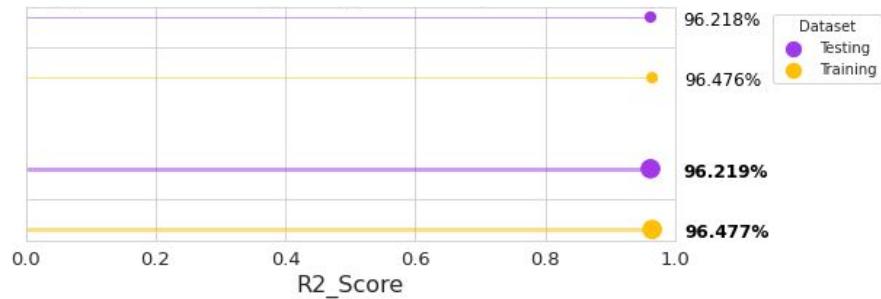
4 data dengan R2 Score tertinggi, dimiliki hampir seluruh dataset kedua

Model: Linear Regression

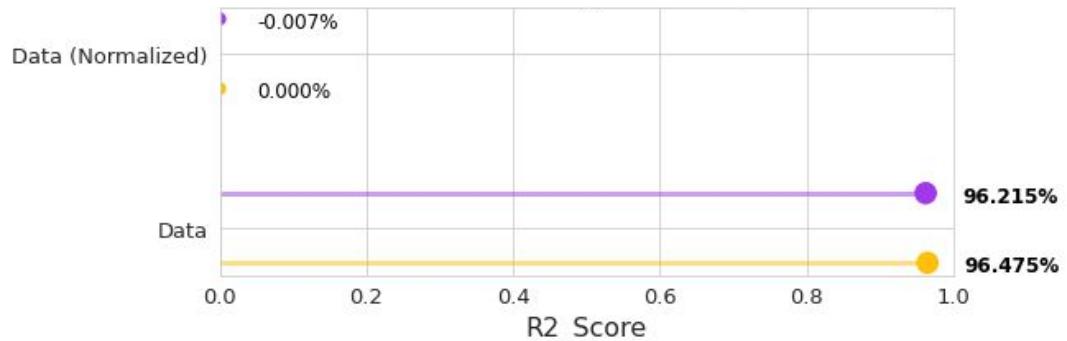
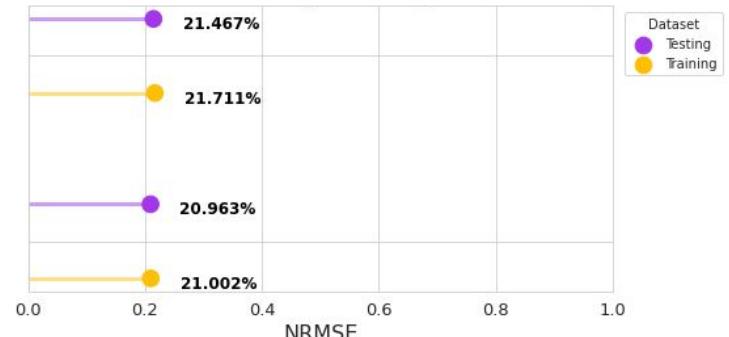
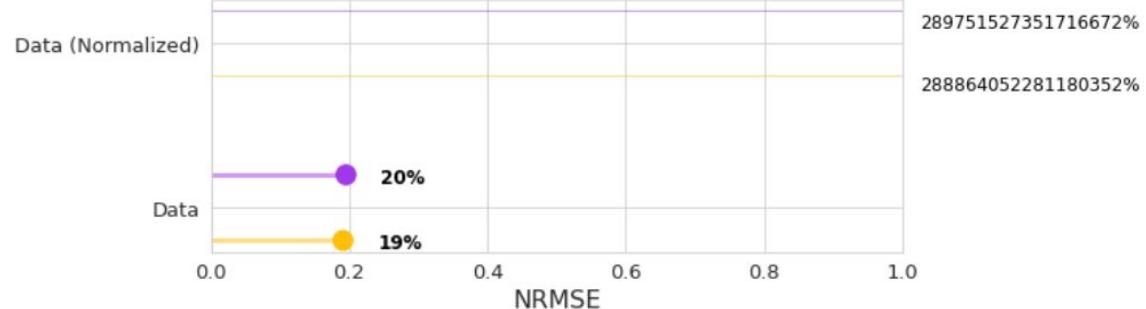
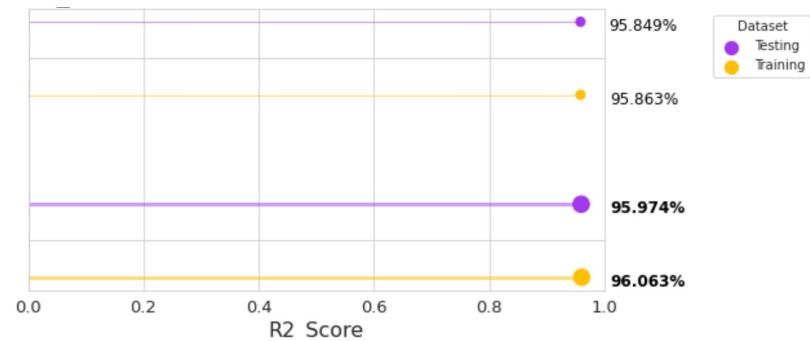


4 data dengan nilai NMRSE terendah dimiliki hampir seluruh dataset kedua

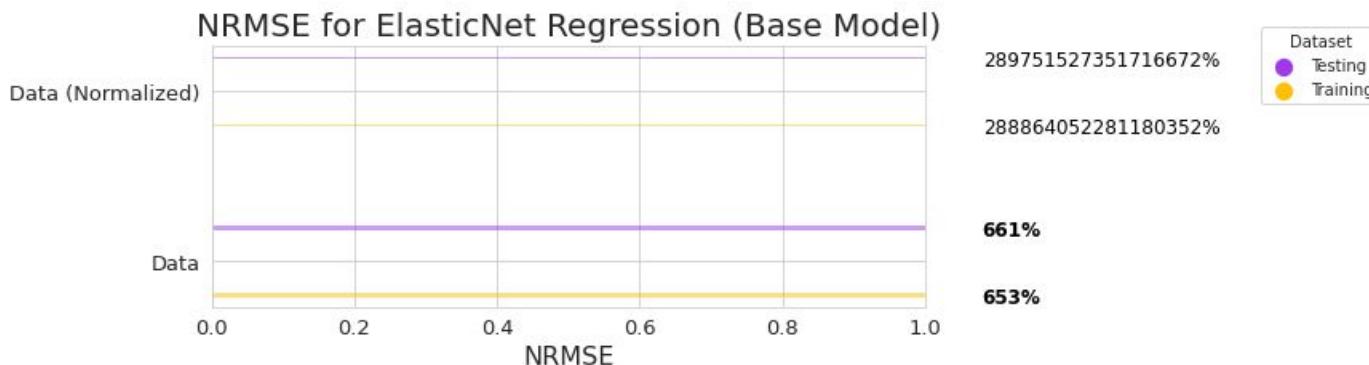
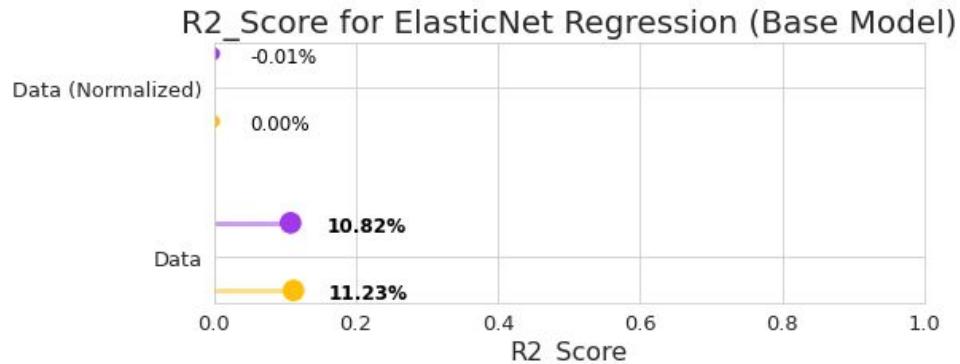
Model: Ridge Regression

Base Model**Tuned Model**

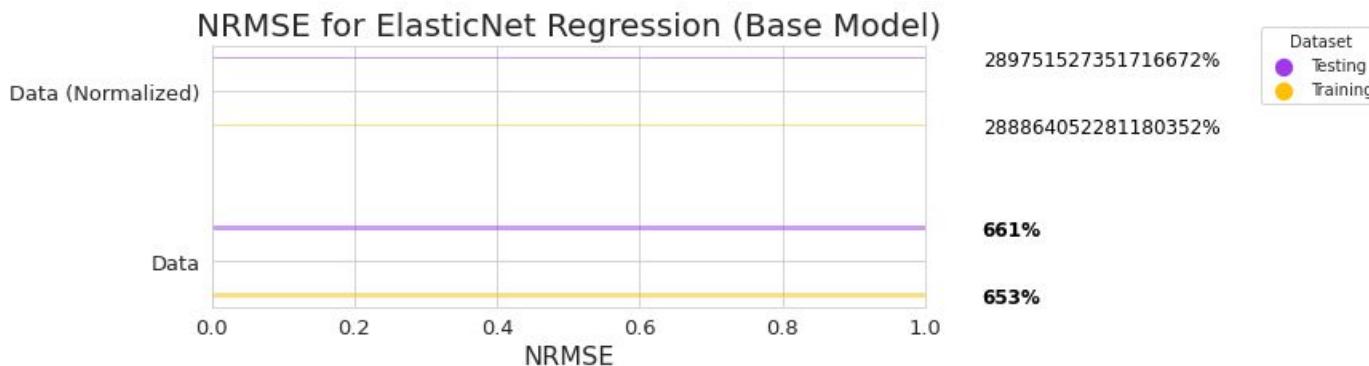
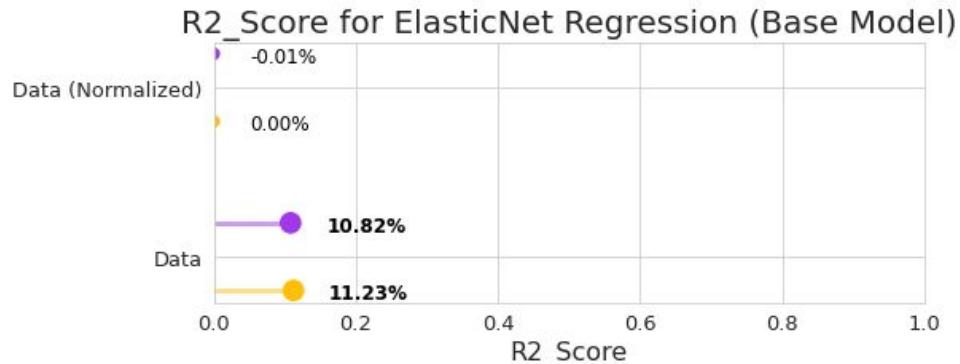
Model: Lasso Regression

Base Model**Tuned Model**

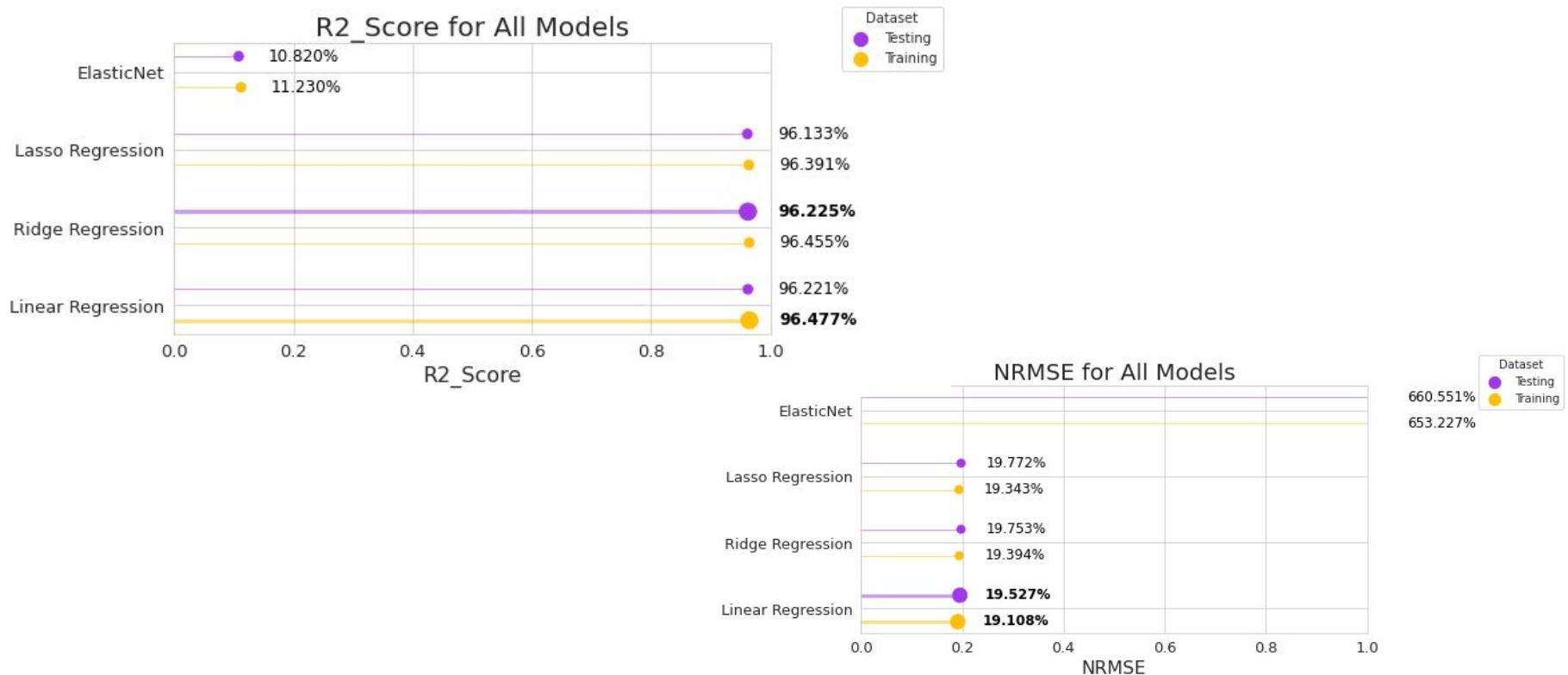
Model: Elastic Net



Model: Elastic Net



Model Comparison



Conclusion

Conclusion

- Model terbaik untuk memprediksi penjualan mingguan adalah Ridge Regression. Model ini dapat memberikan informasi yang membantu manajer bisnis mengidentifikasi dan memahami kelemahan dalam perencanaan bisnis.

Rekomendasi:

- Divisi Marketing sebaiknya meningkatkan advertising ketika Minggu-minggu sebelum Christmas dan saat Thanksgiving.
- Perlu penambahan orang dari divisi Logistik di bulan November-Desember karena penjualannya meningkat signifikan dibandingkan bulan-bulan lainnya.
- Melakukan re-stock barang secukupnya di hari biasa untuk meminimalkan production cost.

Terima kasih!

Ada pertanyaan?

zenius

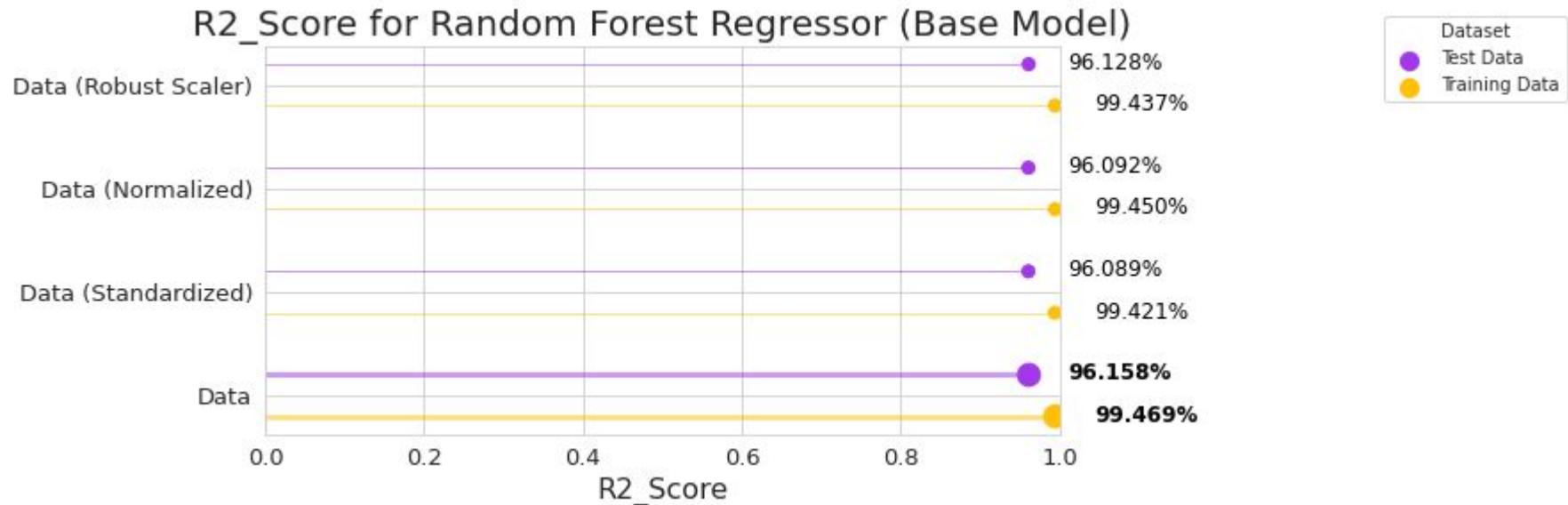


**Kampus
Merdeka**
INDONESIA JAYA

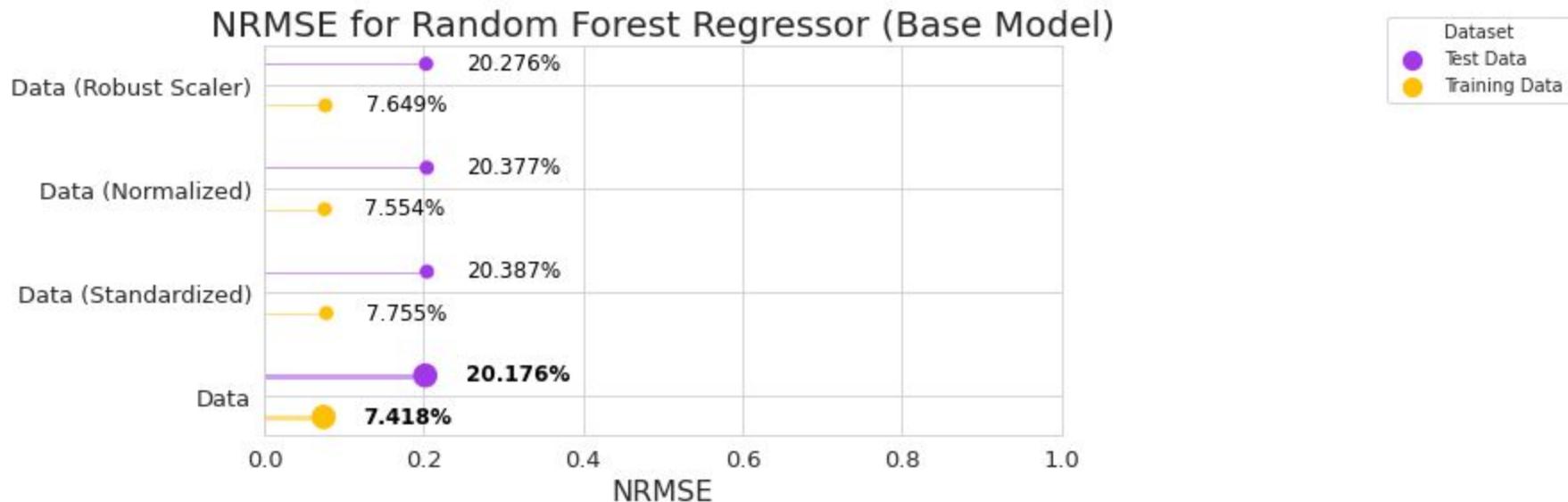


LAMPIRAN

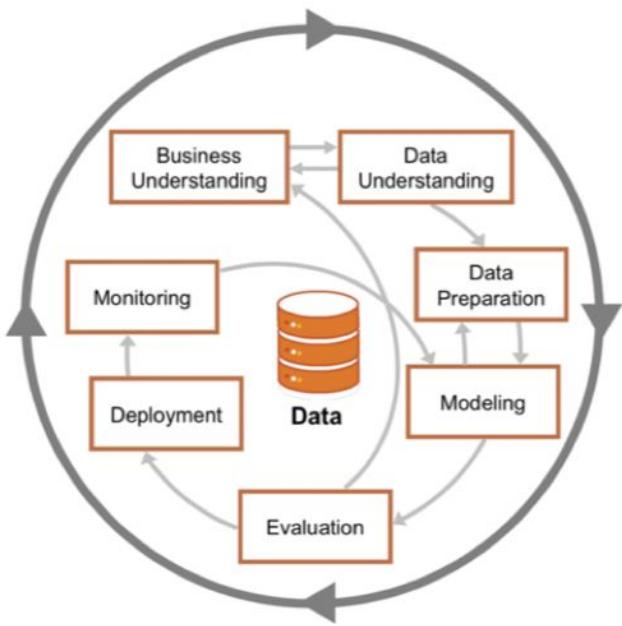
Model: Random Forest Regressor



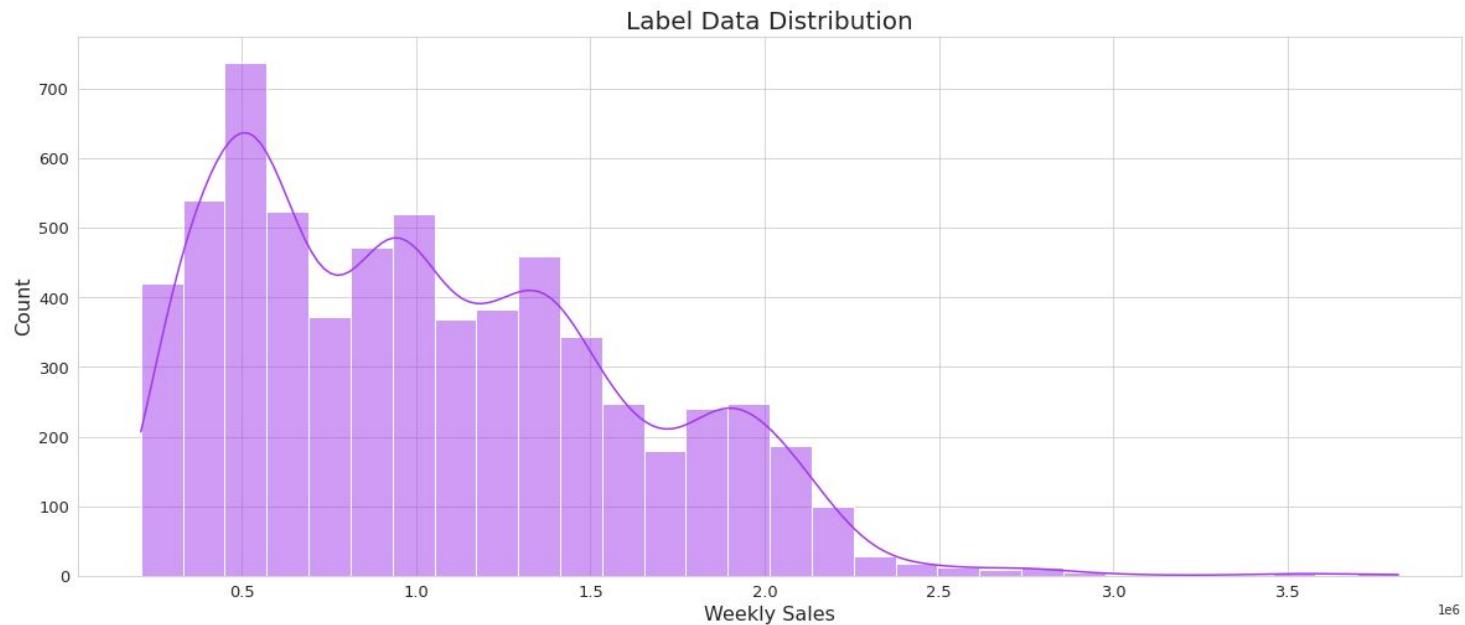
Model: Random Forest Regressor



CRISP-DM Methodology

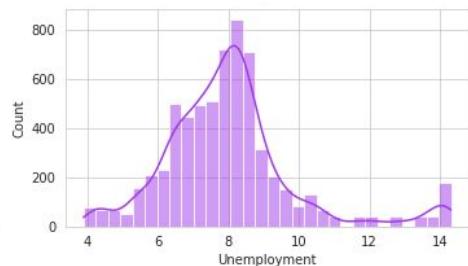
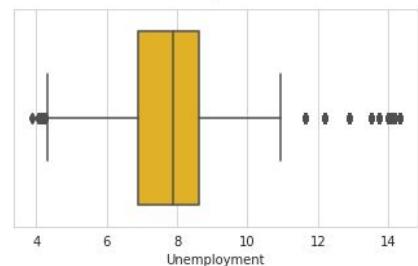
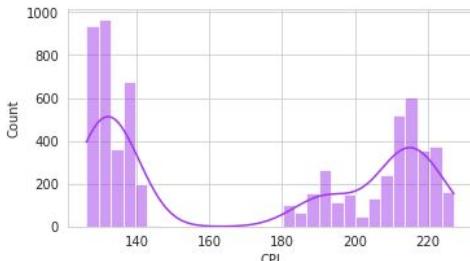
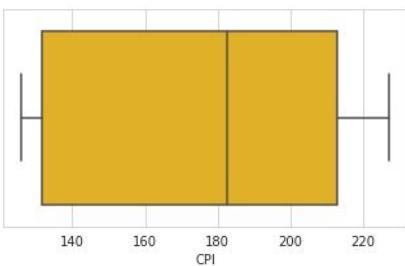
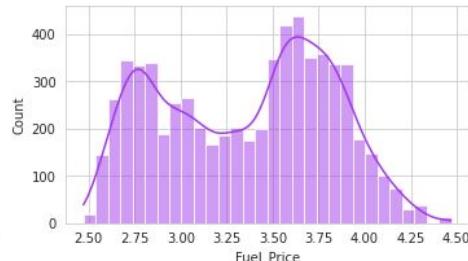
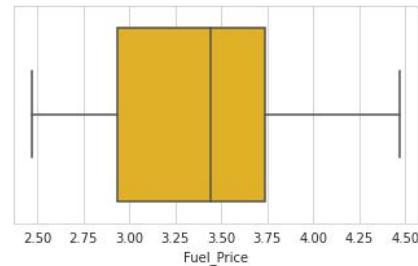
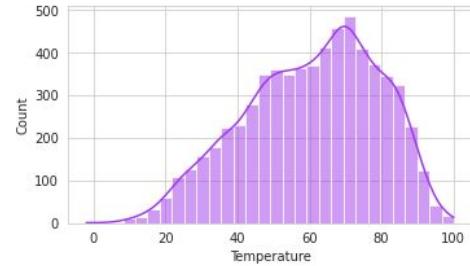
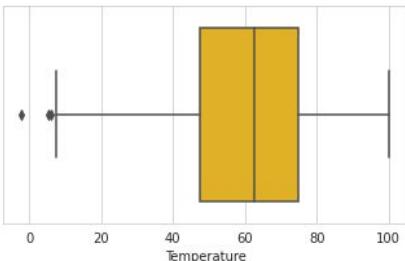


Univariate Analysis (Target)



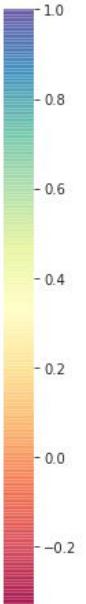
Univariate Analysis (Numerical)

Data Distribution of Numerical Features

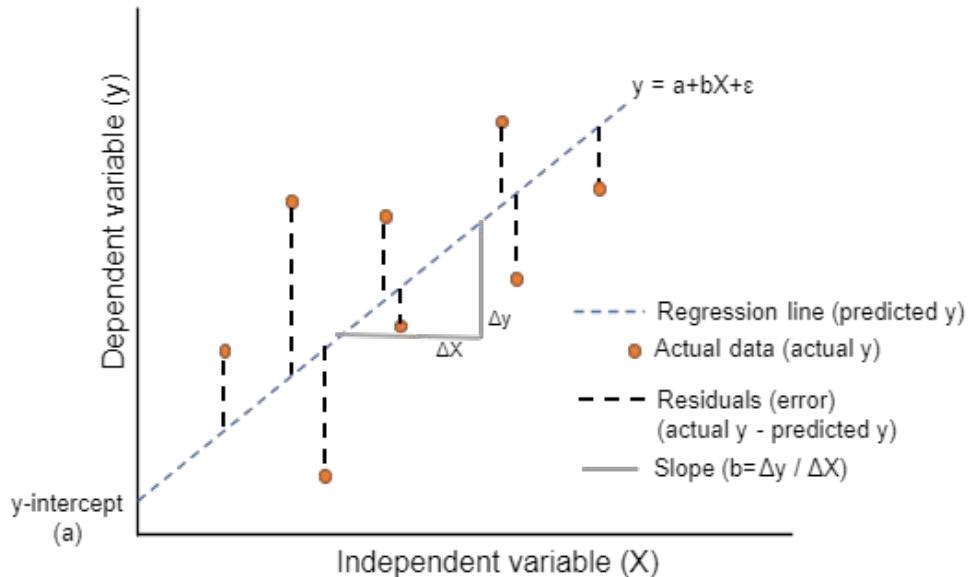


Correlation Matrix

	Store	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment	Day	Month	Year	WeekOfYear	Super_Bowl	Labor_Day	Thanksgiving	Christmas	Before_Super_Bowl	Before_Labor_Day	Before_Thanksgiving	Before_Christmas	Before_Super_Bowl	Before_Labor_Day	Before_Thanksgiving	Before_Christmas	Before_Unemployment_class	CPI_class
Store	1	-0.34	-4.4e-16	0.023	0.06	-0.21	0.22	1.5e-15	2.9e-15	3.5e-12	3.1e-15	6.2e-182	2e-164	1e-193	5e-186	9e-173	1e-162	1e-178	7e-18	0.2	-0.31				
Weekly_Sales	-0.34	1	0.037	-0.064	0.0095	-0.073	-0.11	-0.017	0.076	-0.018	0.074	0.0083	-0.0012	0.09	-0.018	0.0012	-0.0031	-0.0062	0.15	-0.087	-0.08				
Holiday_Flag	4.4e-16	0.037	1	-0.16	-0.078	0.0022	0.011	0.045	0.12	-0.057	0.13	0.53	0.53	0.43	0.43	-0.04	-0.04	-0.033	-0.033	0.014	0.013				
Temperature	-0.023	-0.064	-0.16	1	0.14	0.18	0.1	0.027	0.24	0.064	0.24	-0.2	0.11	-0.077	-0.15	-0.19	0.14	-0.068	-0.13	0.098	0.29				
Fuel_Price	0.06	0.0095	-0.078	0.14	1	-0.17	-0.035	0.028	-0.042	0.78	-0.032	-0.076	0.018	-0.047	-0.052	-0.082	0.0093	-0.041	-0.052	-0.05	0.067				
CPI	-0.21	-0.073	-0.0022	0.18	-0.17	1	-0.3	0.0027	0.005	0.075	0.006	-0.0035	0.002	-0.0018	0.00097	-0.004	0.0018	-0.0019	-0.0012	-0.3	0.69				
Unemployment	0.22	-0.11	0.011	0.1	-0.035	-0.3	1	-0.0042	-0.013	-0.24	-0.016	0.011	-0.0061	0.0089	0.0089	0.011	-0.0061	0.0089	0.0089	0.96	-0.31				
Day	1.5e-15	-0.017	0.045	0.027	0.028	0.0027	-0.0042	1	0.015	0.0064	0.1	-0.078	-0.12	0.13	0.2	-0.2	-0.062	0.038	0.11	-0.0036	0.0043				
Month	2.9e-15	0.076	0.12	0.24	-0.042	0.005	-0.013	0.015	1	-0.19	1	-0.2	0.12	0.17	0.2	-0.2	0.1	0.17	0.2	-0.005	0.0092				
Year	3.5e-12	-0.018	-0.057	0.064	0.78	0.075	-0.24	0.0064	-0.19	1	-0.18	0.0064	0.0064	-0.069	-0.069	0.0064	0.0064	-0.069	-0.069	-0.24	0.25				
WeekOfYear	3.1e-15	0.074	0.13	0.24	-0.032	0.006	-0.016	0.1	1	-0.18	1	-0.21	0.11	0.18	0.22	-0.22	0.095	0.17	0.21	-0.0081	0.012				
Super_Bowl	6.2e-18	0.0083	0.53	-0.2	-0.076	-0.0035	0.011	-0.078	-0.2	0.0064	-0.21	1	-0.021	-0.017	-0.017	-0.021	-0.021	-0.017	-0.017	0.012	-0.015				
Labor_Day	2.2e-16	0.0012	0.53	0.11	0.018	0.002	-0.0061	-0.12	0.12	0.0064	0.11	-0.021	1	-0.017	-0.017	-0.021	-0.021	-0.017	-0.017	-0.0071	0.0071				
Thanksgiving	4.1e-19	0.09	0.43	-0.077	-0.047	-0.0018	0.0089	0.13	0.17	-0.069	0.18	-0.017	-0.017	1	-0.014	-0.017	-0.017	-0.014	-0.014	0.013	-0.0093				
Christmas	8.6e-18	-0.018	0.43	-0.15	-0.052	-0.00097	0.0089	0.2	0.2	-0.069	0.22	-0.017	-0.017	-0.014	1	-0.017	-0.017	-0.014	-0.014	0.013	-0.0093				
Before_Super_Bowl	6.9e-17	0.0012	-0.04	-0.19	-0.082	-0.004	0.011	-0.2	-0.2	0.0064	-0.22	0.021	-0.021	-0.017	-0.017	1	-0.021	-0.017	-0.017	0.012	-0.015				
Before_Labor_Day	3.1e-16	0.0031	-0.04	0.14	0.0093	0.0018	-0.0061	-0.062	0.1	0.0064	0.095	-0.021	-0.021	-0.017	-0.017	-0.021	1	-0.017	-0.017	0.0071	0.0071				
Before_Thanksgiving	2.1e-17	0.0062	-0.033	-0.068	-0.041	-0.0019	0.0089	0.038	0.17	-0.069	0.17	0.017	0.017	-0.014	-0.014	-0.017	-0.017	1	-0.014	0.013	-0.0093				
Before_Christmas	8.7e-18	0.15	-0.033	-0.13	-0.052	-0.0012	0.0089	0.11	0.2	-0.069	0.21	0.017	0.017	-0.014	-0.014	-0.017	-0.017	-0.014	1	0.013	-0.0093				
Unemployment_class	0.2	-0.087	0.014	0.098	-0.05	-0.3	0.96	-0.0036	-0.005	-0.24	0.0081	0.012	-0.0071	0.013	0.013	0.012	-0.0071	0.013	0.013	1	-0.28				
CPI_class	-0.31	-0.08	-0.013	0.29	0.067	0.69	-0.31	0.0043	0.0092	0.25	0.012	0.015	0.0071	-0.0093	-0.0093	-0.015	0.0071	-0.0093	-0.0093	-0.28	1				



Linear Regression Explained



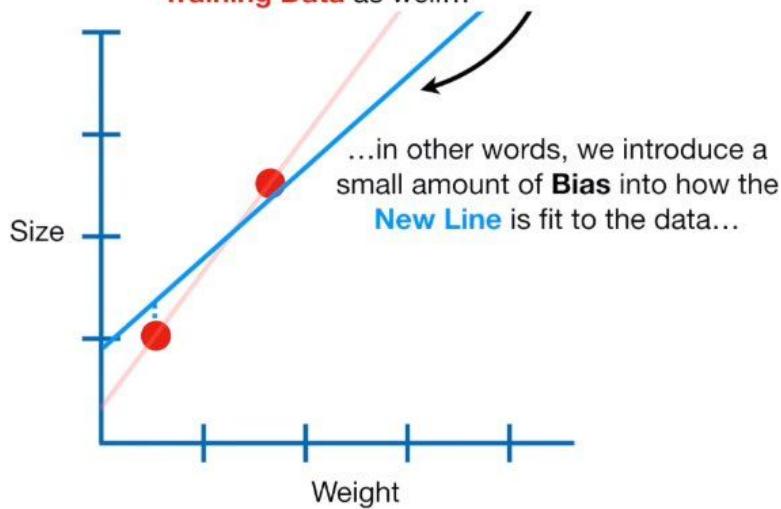
$$y_i = b_0 + b_1 x + e$$

Annotations pointing to the equation:

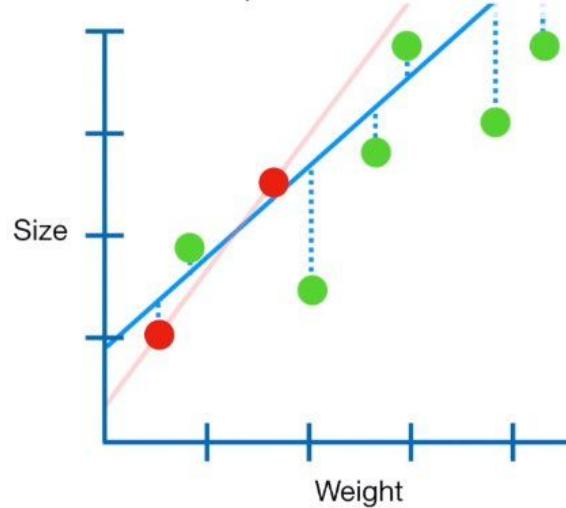
- Estimated (or predicted) y value
- Estimate of the regression intercept
- Estimate of the regression slope
- Independent variable
- Error term

Ridge Regression Explained

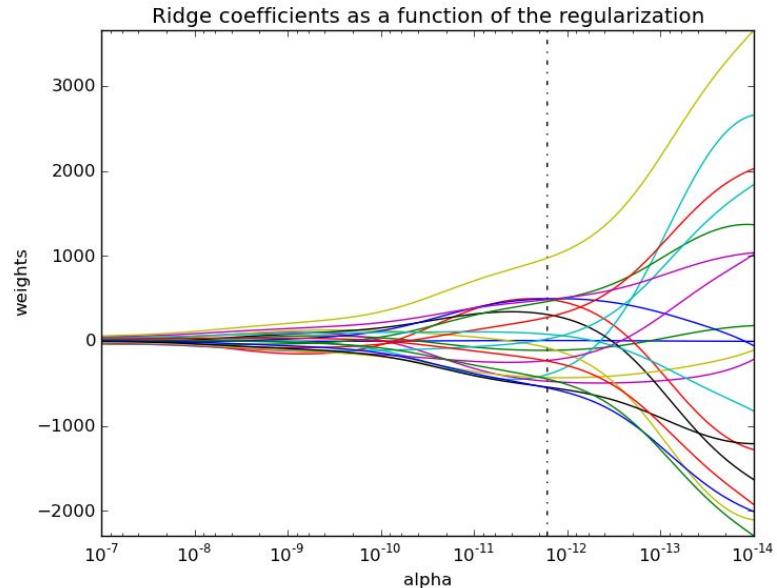
The main idea behind **Ridge Regression** is to find a **New Line** that doesn't fit the **Training Data** as well...



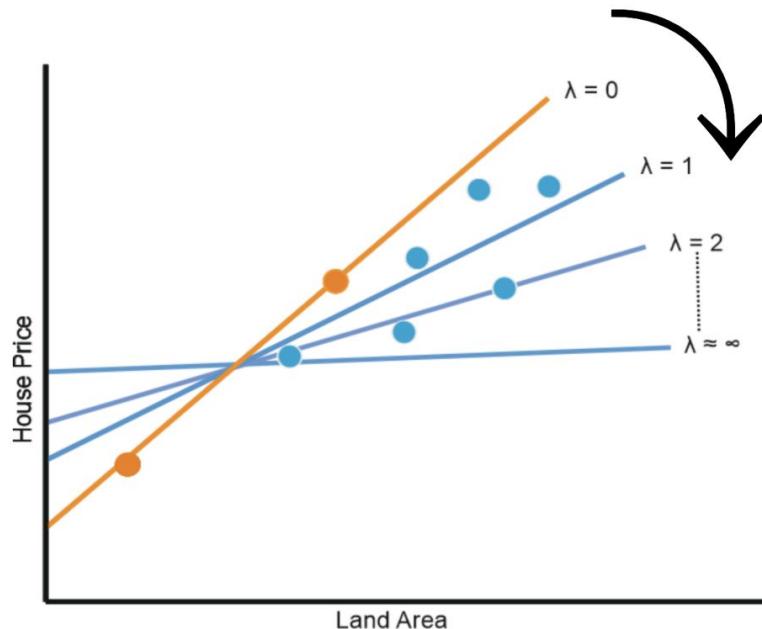
In other words, by starting with a slightly worse fit, **Ridge Regression** can provide better long term predictions.



Ridge Regression Explained



Lasso Explained



Lasso Explained

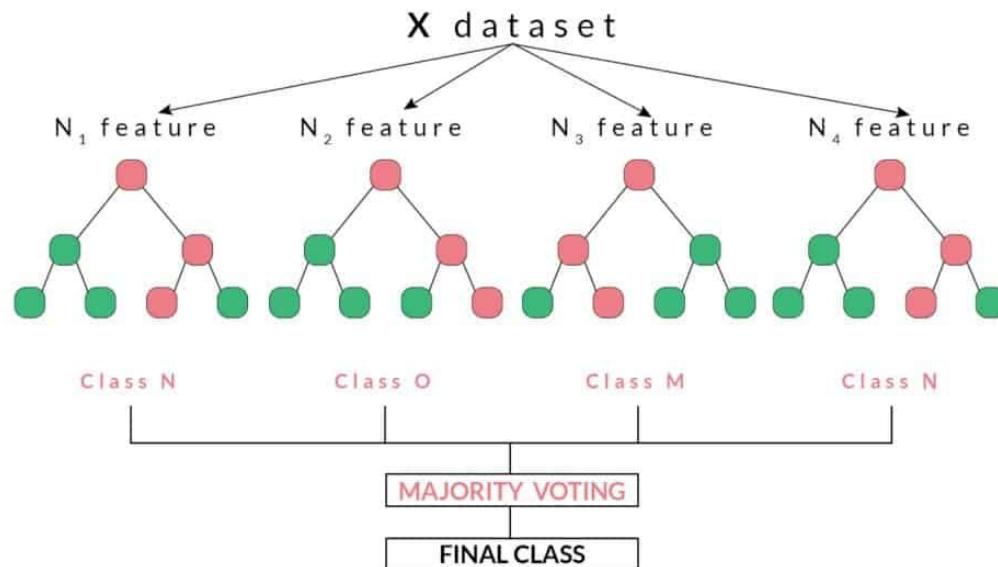
Base model dari Lasso memiliki **performance** yang bisa dibilang buruk. Dataset yang kita olah memiliki banyak **categorical predictor** seperti Store, Holiday_Flag, Month, Week, WeekOfYear, Day, dan yang lainnya. Berdasarkan hasil **feature importance** yang didapatkan dari koefisien regresor, pada dataset ini **numerical predictor** justru memiliki nilai **importance** yang sangat rendah dibandingkan **categorical predictor**.

Algoritma Lasso dari ScikitLearn sepertinya tidak bisa menangani masalah ini karena Lasso memberikan nilai **penalty** terhadap fitur yang memiliki kontribusi minor sehingga koefisiennya mendekati/menjadi nol [\[1\]](#), sedangkan untuk **categorical predictor** yang telah di-encode akan terpecah menjadi beberapa variabel (Contoh: fitur "Store" akan menjadi "Store_1", "Store_2", "Store N") dan tidak masuk akal untuk mengambil sebagian variabel saja dari **categorical predictor**. Misalnya, setelah diketahui nilai koefisiennya, "Store_27" dan "Store_31" tidak diberikan penalti karena memiliki kontribusi lebih besar, tetapi "Store _44" memiliki kontribusinya yang sangat kecil sehingga Lasso memberi penalti dengan menjadikan koefisiennya 0, padahal tidak bisa begitu karena mereka merupakan satu kesatuan dan seharusnya seluruh fitur-fitur yang telah di-encode dari fitur "Store" diambil semua atau tidak dipakai semua) sehingga diperlukan **treatment** khusus (modifikasi Lasso) untuk masalah seperti ini. [\[2\]](#) [\[3\]](#). Sementara, untuk alasan mengapa Lasso memiliki **performance** yang baik pada dataset tanpa **feature scaling** belum kami temukan karena berdasarkan sumber yang ada Lasso seharusnya memiliki **performance** yang baik jika skala datanya sama [\[\[4\]\]](#)(<https://stats.stackexchange.com/questions/86434/is-standardisation-before-lasso-really-necessary>).

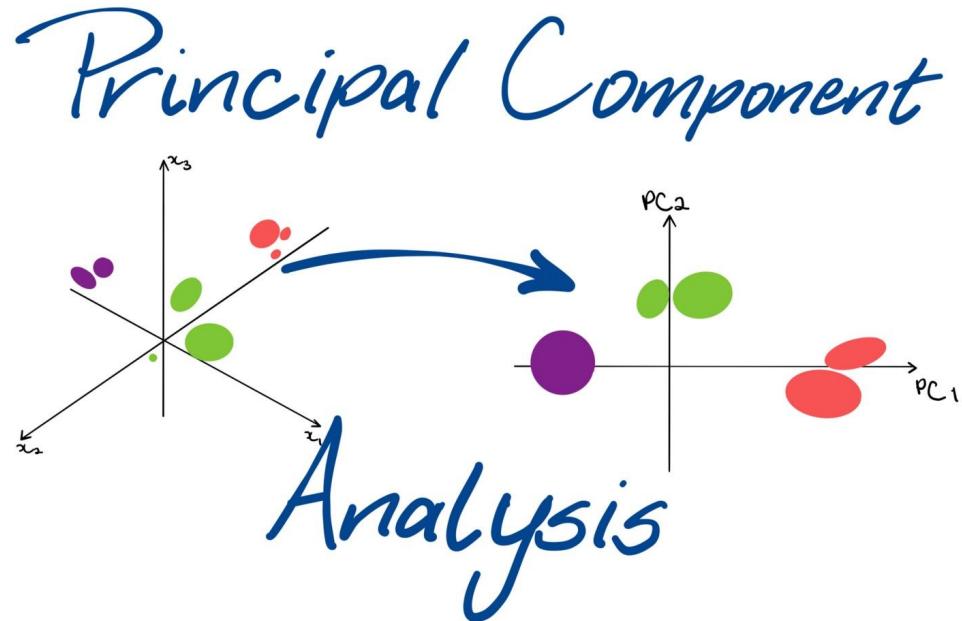
Lasso Explained

Sementara setelah dilakukan hyperparameter tuning, dengan alpha yang mendekati 0, *performance* model langsung meningkat karena semakin kecil alpha maka semakin kecil penalti yang diberikan oleh Lasso. Berdasarkan dokumentasi dari Scikit Learn [[5]](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html), jika nilai alpha di-set mendekati atau sama dengan nol, maka cara kerja Lasso secara esensinya sama saja dengan Linear Regression standard karena efek penalti yang diberikan hampir tidak ada. Oleh karena itu saat tuning dengan alpha mendekati 0, *performance* model baik nilai R2 score dan NRMSE-nya langsung berubah drastis.

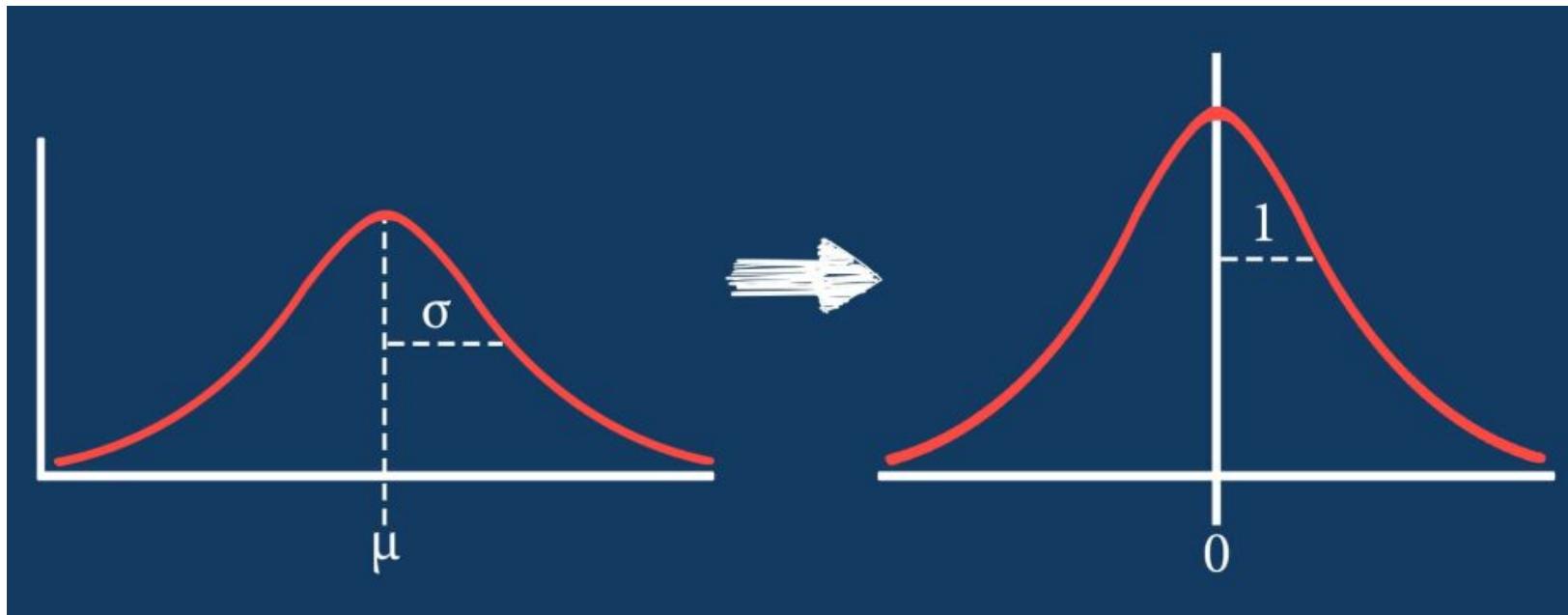
Random Forest Explained



Principal Component Analysis



Standardization



Evaluation Metric: MAE (Mean Absolute Error)

$$MAE = \frac{1}{n} \sum \text{Sum of } |y - \hat{y}|$$

Divide by the total number of data points

Predicted output value

Actual output value

The absolute value of the residual

Evaluation Metric: MSE (Mean Squared Error)

$$MSE = \frac{1}{n} \sum \underbrace{\left(y - \hat{y} \right)^2}_{\text{The square of the difference between actual and predicted}}$$

The square of the difference
between actual and
predicted

Evaluation Metric: RMSE (Root Mean Squared Error)

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (Predicted_i - Actual_i)^2}{N}}$$

Evaluation Metric: MAPE (Mean Absolute Percentage Error)

$$MAPE = \frac{100\%}{n} \sum \left| \frac{\hat{y} - y}{y} \right|$$

Multiplying by 100% converts to percentage

The residual

Each residual is scaled against the actual value

Evaluation Metric: NRMSE (Normalized Root Mean Squared Error)

$$NRMSE = \frac{RMSE}{mean(y)} \text{ OR } NRMSE = \frac{RMSE}{y_{max} - y_{min}}$$

Evaluation Metric: R2-Score

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2}$$

Feature Selection & Decomposition

- Recursive Feature Elimination (RFE)
menyeleksi fitur berdasarkan estimator secara rekursif
- Principal Component Analysis (PCA)
mereduksi fitur dengan melakukan kombinasi berdasarkan nilai varians