CHRISTIAN

Customer Lifetime Value Prediction

Content

- Part 1: Business Problem Understanding
- Part 2: Data Understanding
- Part 3: Data Preprocessing
- Part 4: Data Preparation
- Part 5: Model & Evaluation
- Part 6: Conclusion & Recommendation

Part 1:

Business Problem Understanding

What Is This Project About?

This project focuses on predicting Customer Lifetime Value (CLV) using historical customer data

CLV represents the total revenue a business expects from a customer over the entire relationship

Accurately predicting CLV helps businesses focus marketing efforts, improve retention, and maximize profit

Why Is CLV Important?

Understand which customers generate the most long-term value

Allocate marketing budgets more effectively

Reduce churn by identifying high-value customers early

Drive decision-making across sales, customer support, and loyalty programs

Project Objective

Build a predictive model to estimate Customer Lifetime Value (CLV)

Determine which machine learning model performs best

Evaluate how accurate the predictions are

Identify which customer features most influence CLV

Part 2:

Data Understanding

Dataset (5669 Rows)

Vehicle Class	Coverage	Renew Offer Type	Employment Status	Marital Status	Education	Number of Policies	Monthly Premium Auto	Total Claim Amount	Income	Customer Lifetime Value

Dataset (5669 Rows)

Vehicle Class	Coverage	Renew Offer Type	Employment Status	Marital Status	Education	Number of Policies	Monthly Premium Auto	Total Claim Amount	Income	Customer Lifetime Value

Category

Numeric

Part 3:

Data Preprocessing

Missing Value

Vehicle Class	Coverage	Renew Offer Type	Employment Status	Marital Status	Education	Number of Policies	Monthly Premium Auto	Total Claim Amount	Income	Customer Lifetime Value
0	0	0	0	0	0	0	0	0	0	0

No missing value were found

Duplicates

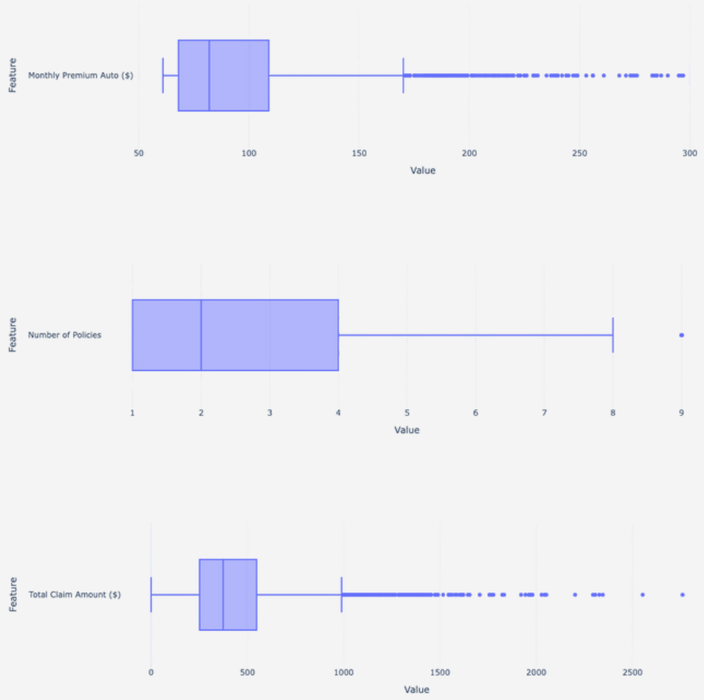
Duplicates	No Duplicates
618	5051

Drop Duplicates

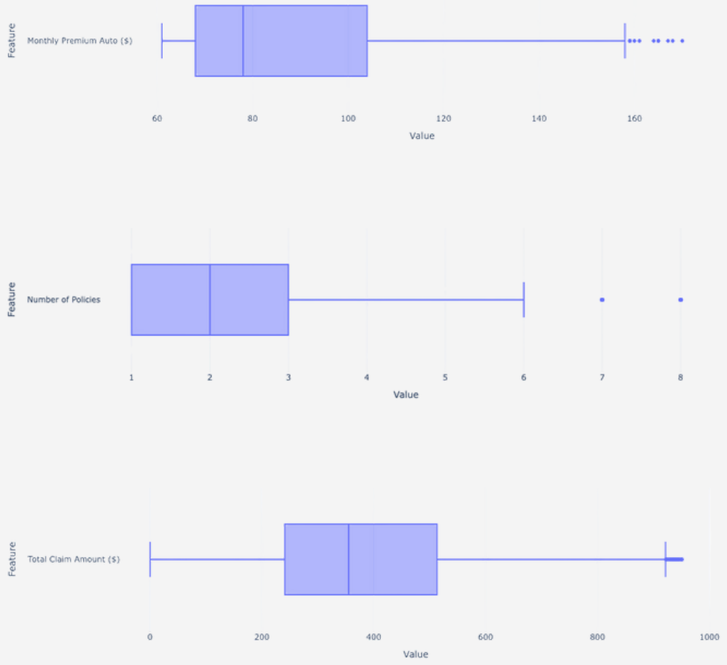
Duplicates	No Duplicates
0	5051

Found and removed 618 duplicate rows from the dataset (≈10.91% of total data).

Outlier



Applied IQR method to Monthly Premium Auto, Total Claim Amount, Number of Policies features



Found and removed 594 outlier rows from the dataset (≈11.76% of total data).

Invalid Value

Vehicle Class	Four-Door Car, Two-Door Car, SUV, Sports Car
Coverage	Extended, Basic, Premium
Renew Offer Type	Offer1, Offer3, Offer2, Offer4
Employment Status	Retired, Employed, Disabled, Medical Leave, Unemployed
Marital Status	Divorced, Married, Single
Education	High School or Below, College, Master, Bachelor, Doctor

	Min	Mean	Max
Number of Policies	1	2.69	8
Monthly Premium Auto	61	86.12	170
Total Claim Amount	0.42	377.75	950.4
Income	0	38610.87	99934
Customer Lifetime Value	1898	7556.8	57520.5

No invalid values were found. All numeric and category fields contain plausible and consistent data.

Part 4:

Data Preparation

Feature and Target

Vehicle Class	Coverage	Renew Offer Type	Employment Status	Marital Status	Education	Number of Policies	Monthly Premium Auto	Total Claim Amount	Income	Customer Lifetime Value

Feature and Target

Vehicle Class	Coverage	Renew Offer Type	Employment Status	Marital Status	Education	Number of Policies	Monthly Premium Auto	Total Claim Amount	Income	Customer Lifetime Value

Category

Target

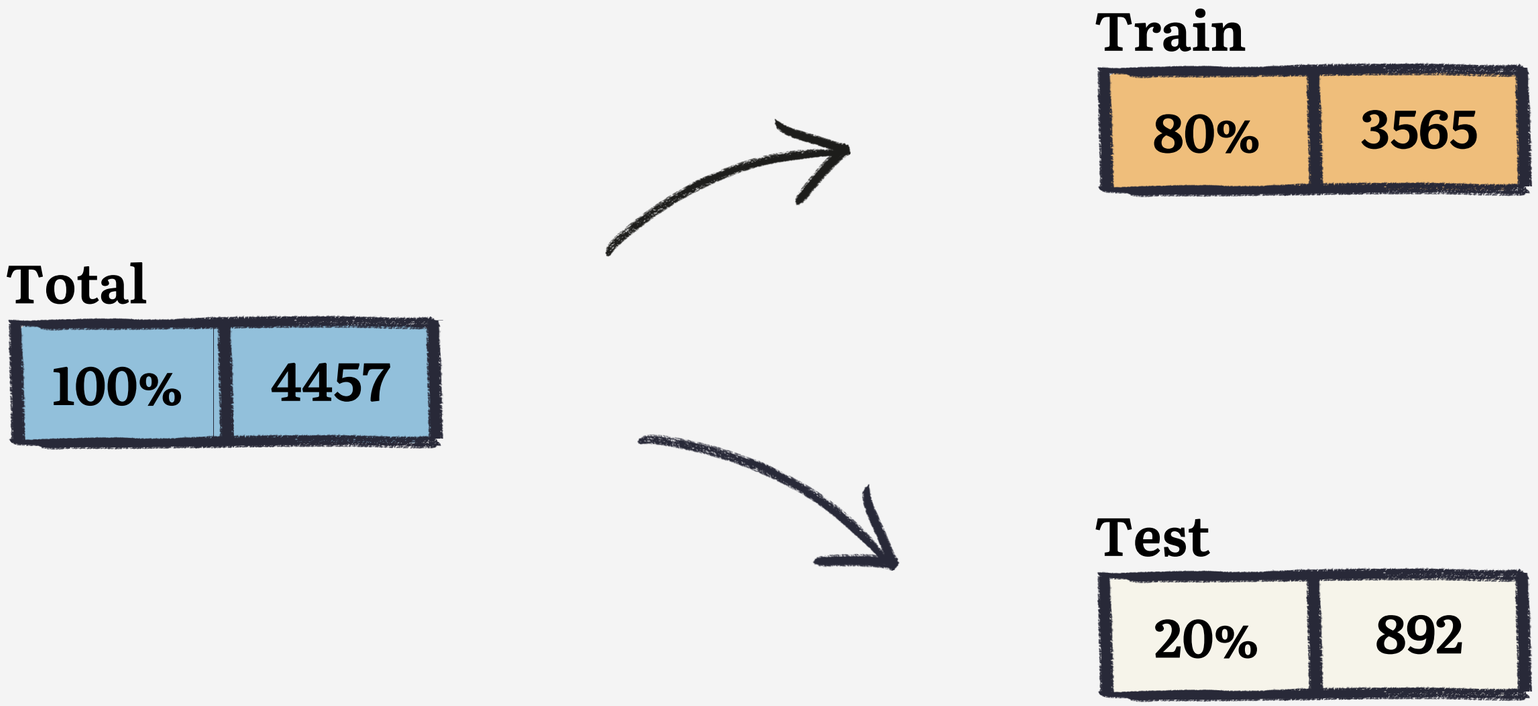
Preprocessing the Data

Vehicle Class	Coverage	Renew Offer Type	Employment Status	Marital Status	Education	Number of Policies	Monthly Premium Auto	Total Claim Amount	Income

One Hot Encoding

Standard Scaller

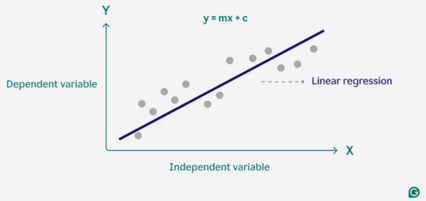
Splitting the Dataset



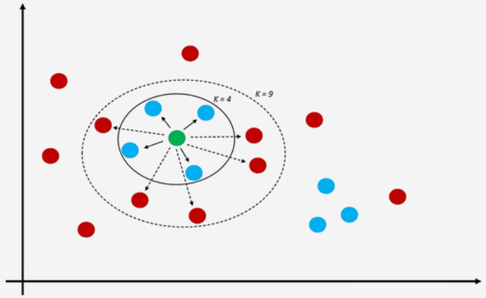
Part 5:

Model & Evaluation

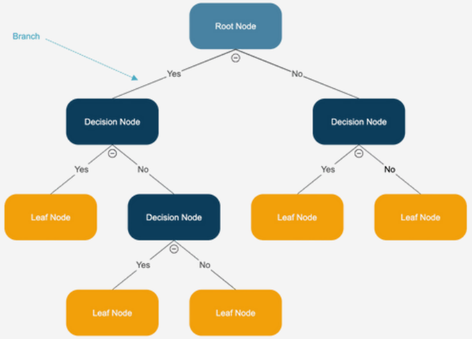
Model Training Overview



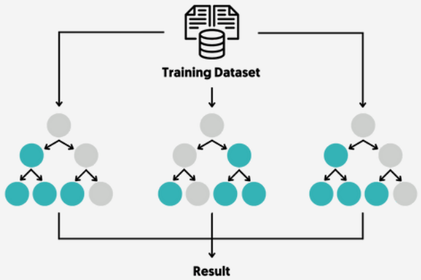
Linear Regression



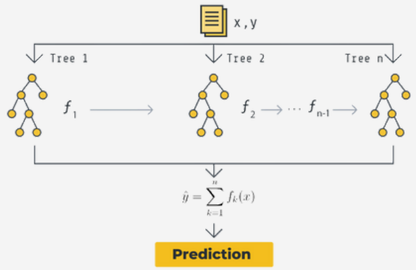
K Nearest Neighbors



Decision Tree

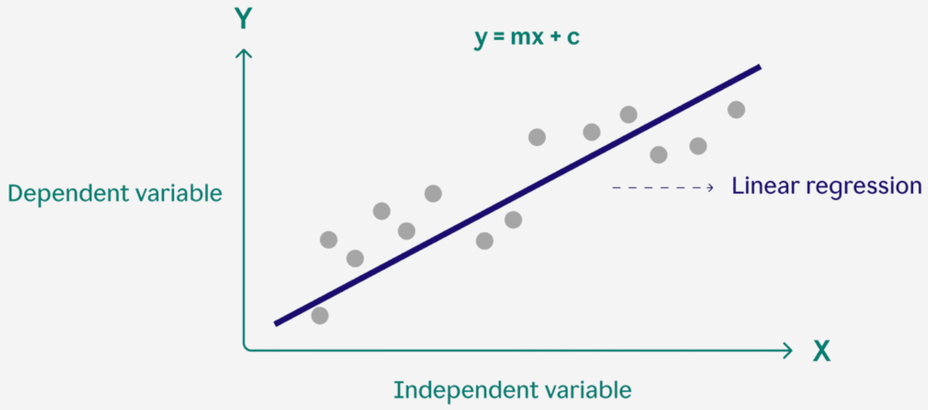


Random Forest



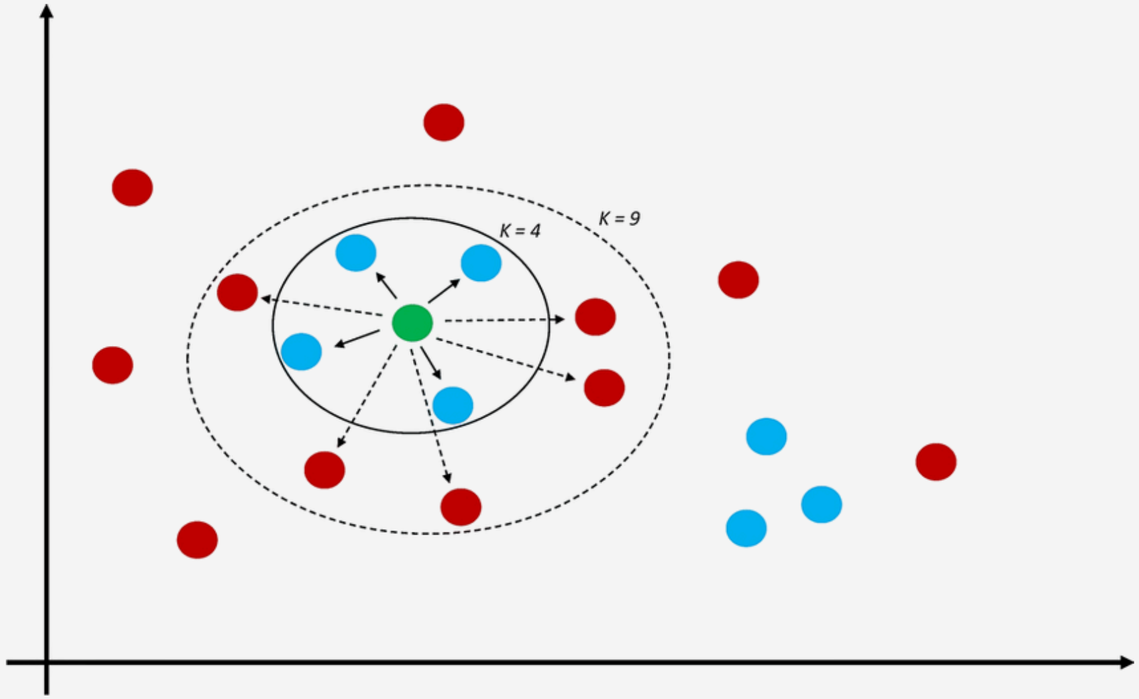
XGBoost

Model Training Overview

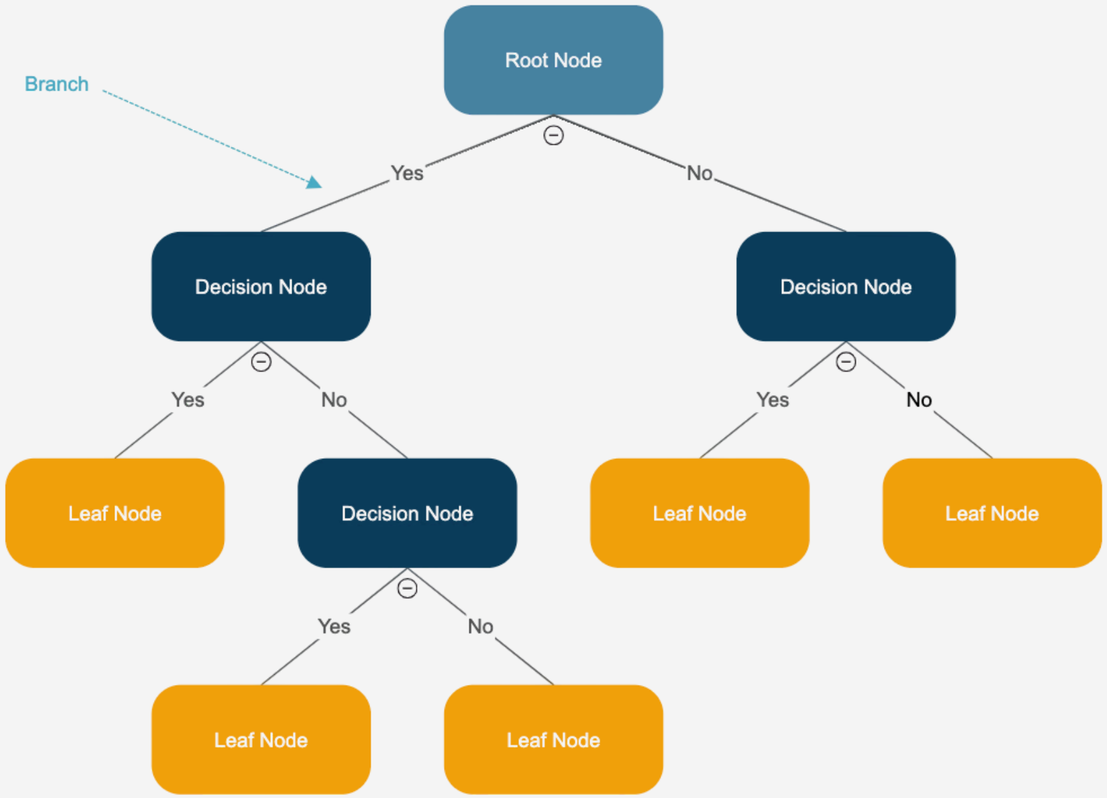


Linear Regression

Overview



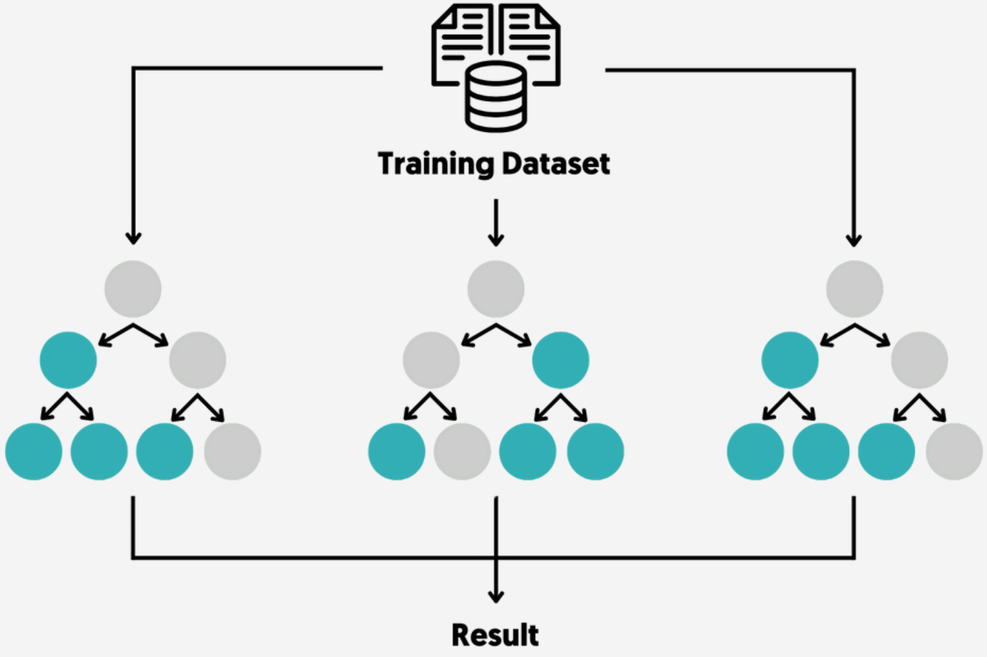
K Nearest Neighbors



Decision Tree

Linear Regression

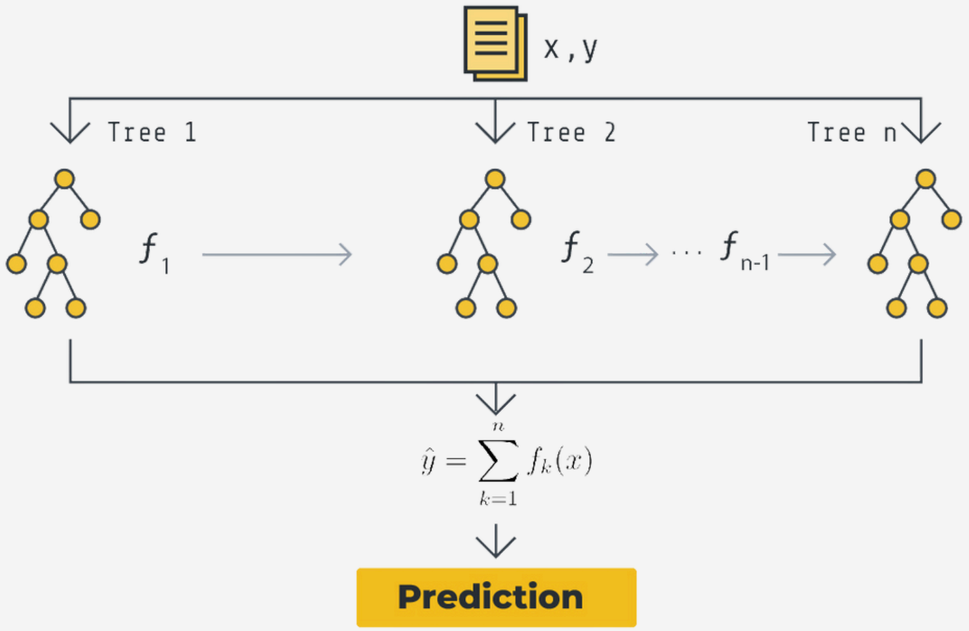
K Nearest Neighbors



Random Forest

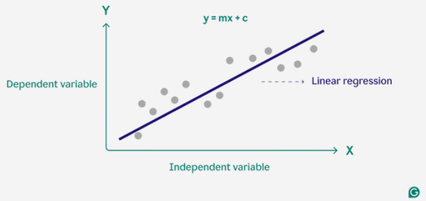
K Nearest Neighbors

Decision Tre

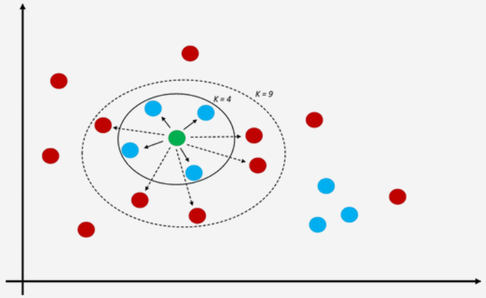


XGBoost

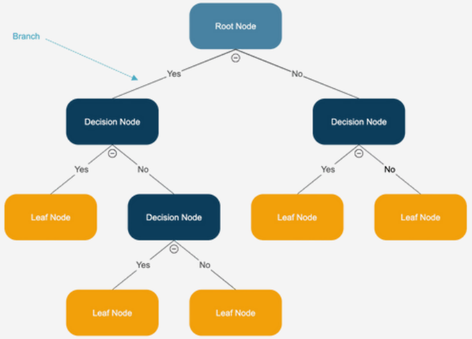
Model Training Overview



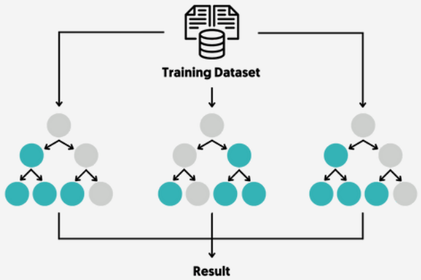
Linear Regression



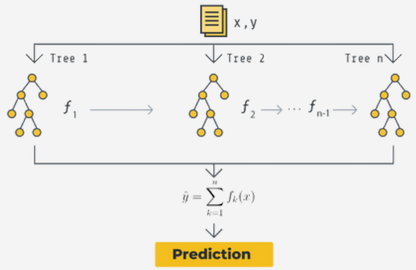
K Nearest Neighbors



Decision Tree



Random Forest



XGBoost

Measured Performance

Mean Absolute Error (MAE)	Mean Absolute Percentage Error (MAPE)	Root Mean Squared Error (RMSE)	R^2 Score

Model Performance Comparison

Model	MAE ($)	MAPE (%)	RMSE ($)	R^2
Linear Regression	3,680.58	61.62	31,046,541.07	0.0919
KNN	3353.25	53.95	30,414,960.28	0.1104
Decision Tree	1718.38	13.64	23,033,690.67	0.3263
Random Forest	1396.73	10.94	11,544,987.68	0.6623
XGBoost	1847.61	21.47	12,494,379.32	0.6346

Model Performance Comparison

Model	MAE ($)	MAPE (%)	RMSE ($)	R^2
Linear Regression	3,680.58	61.62	31,046,541.07	0.0919
KNN	3353.25	53.95	30,414,960.28	0.1104
Decision Tree	1718.38	13.64	23,033,690.67	0.3263
Random Forest	1396.73	10.94	11,544,987.68	0.6623
XGBoost	1847.61	21.47	12,494,379.32	0.6346

Tuning Random Forest with GridSearchCV

max_depth: None
max_features: sqrt
min_samples_leaf: 2
min_samples_split: 5
n_estimators: 200

MAE ($)	MAPE (%)	RMSE ($)	R^2
1,851.24	20.00	13,107,426.99	0.6166

Tuning Random Forest with RandomizedSearchCV

bootstrap: False
max_depth: 14
max_features: log2
min_samples_leaf: 2
min_samples_split: 7
n_estimators: 910

MAE ($)	MAPE (%)	RMSE ($)	R^2
1,813.59	19.37	12,790,503.51	0.6259

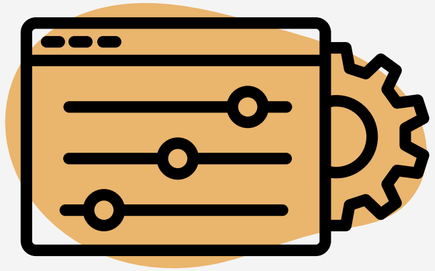
Part 6:

Conclusion & Recommendation

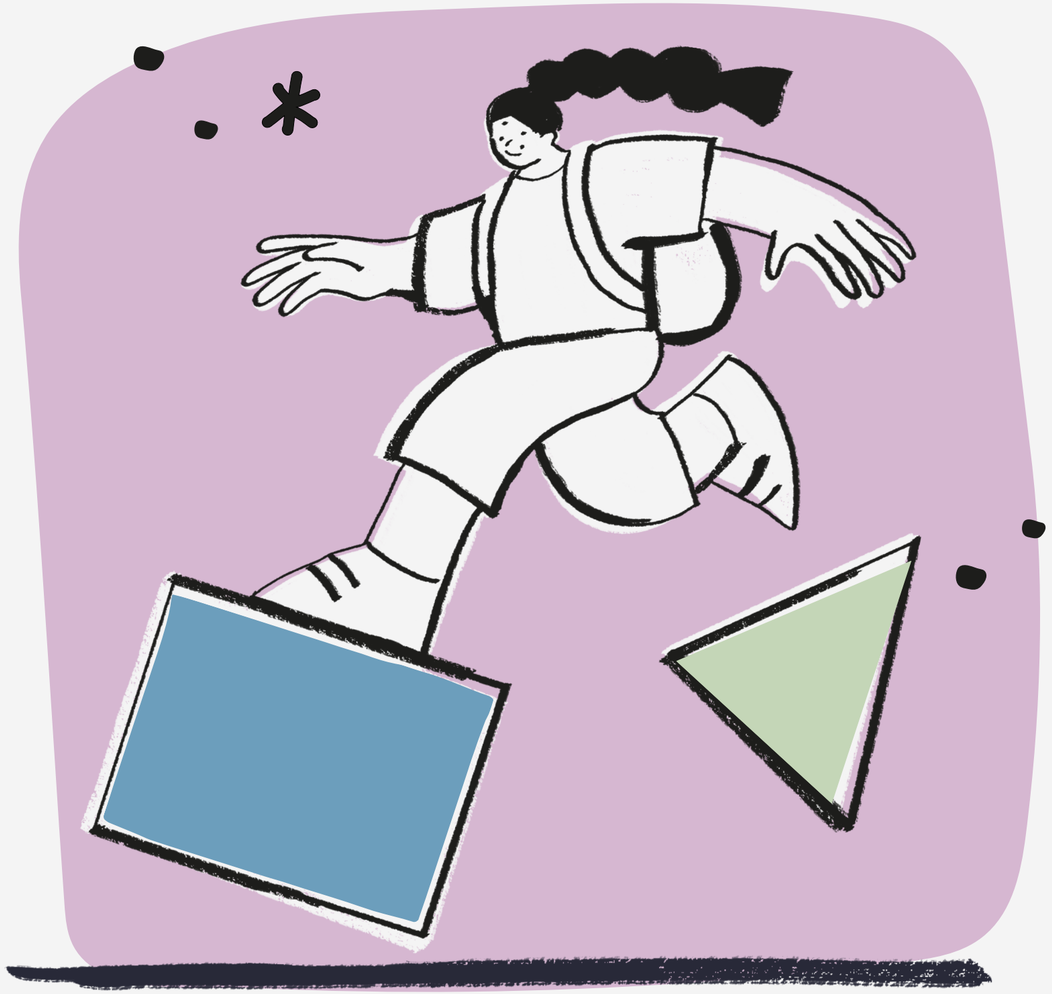
Conclusion:

RF

Random Forest Performed Best



Tuning Didn't Help

Recommendation:

- Use the baseline Random Forest model without tuning as it already performs optimally.
- Retrain the model periodically to maintain performance.
- Collect additional data