Predicting Insurance Claim Severity

Anant Kumar Verma

Introduction

Business Problem

- Insurance companies need to predict claim severity to optimize pricing and risk assessment.
- High claim amounts can impact profitability, while underpricing can lead to losses.

Objective

- Build a predictive model to estimate claim amounts based on vehicle, driver, and policy attributes.
- Use machine learning techniques to identify key factors influencing claim severity.
- Provide actionable insights for improving risk management and pricing strategies.

Dataset Overview

Key Variables in the Dataset

- ❑ **Policy-related:** Policy age, claims history
- ❑ **Vehicle-related:** Age of car, fuel type, engine power, safety features
- ❑ **Driver-related:** Age of policyholder, driving history
- ❑ **Geographical & Environmental Factors:** Area, population density

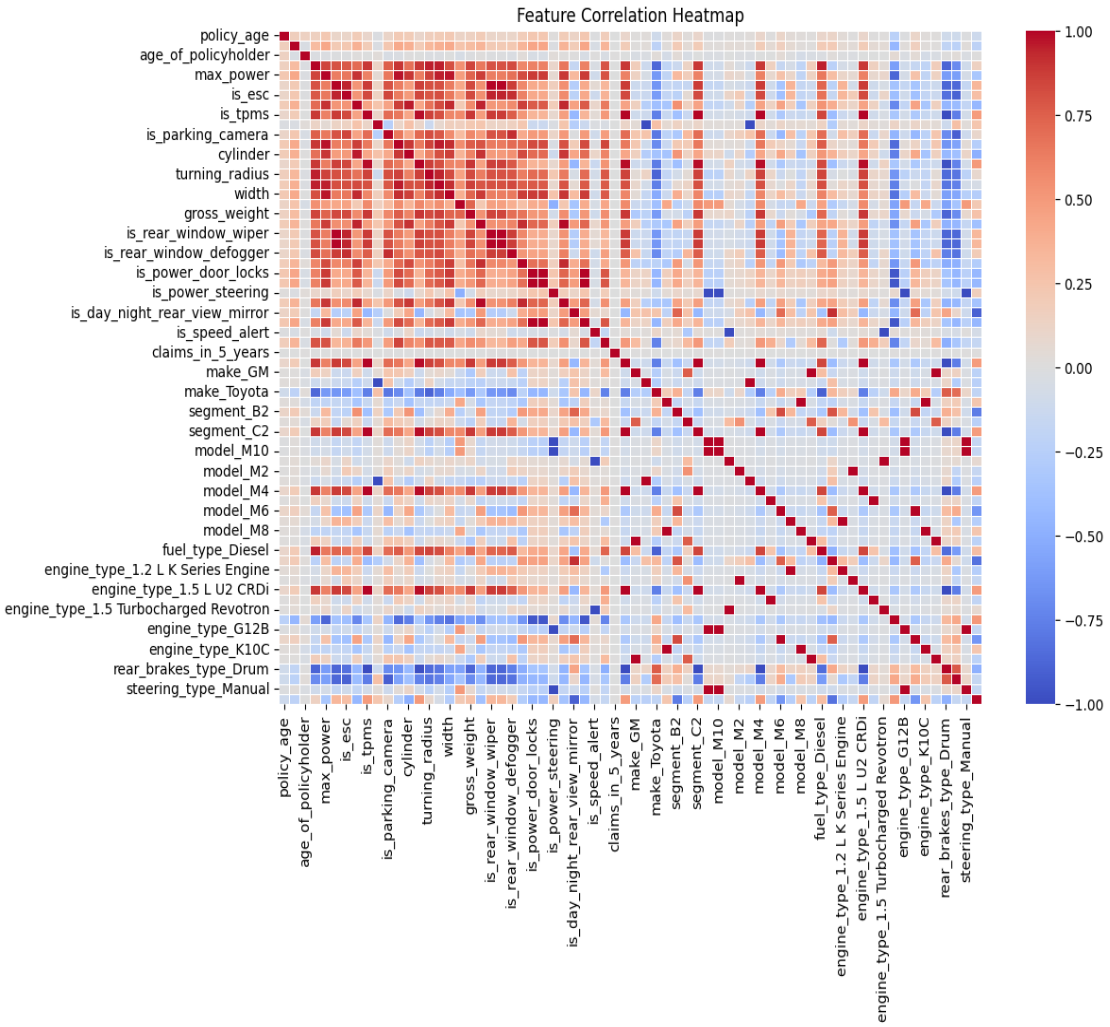
Source: kaggle

Exploratory Data Analysis

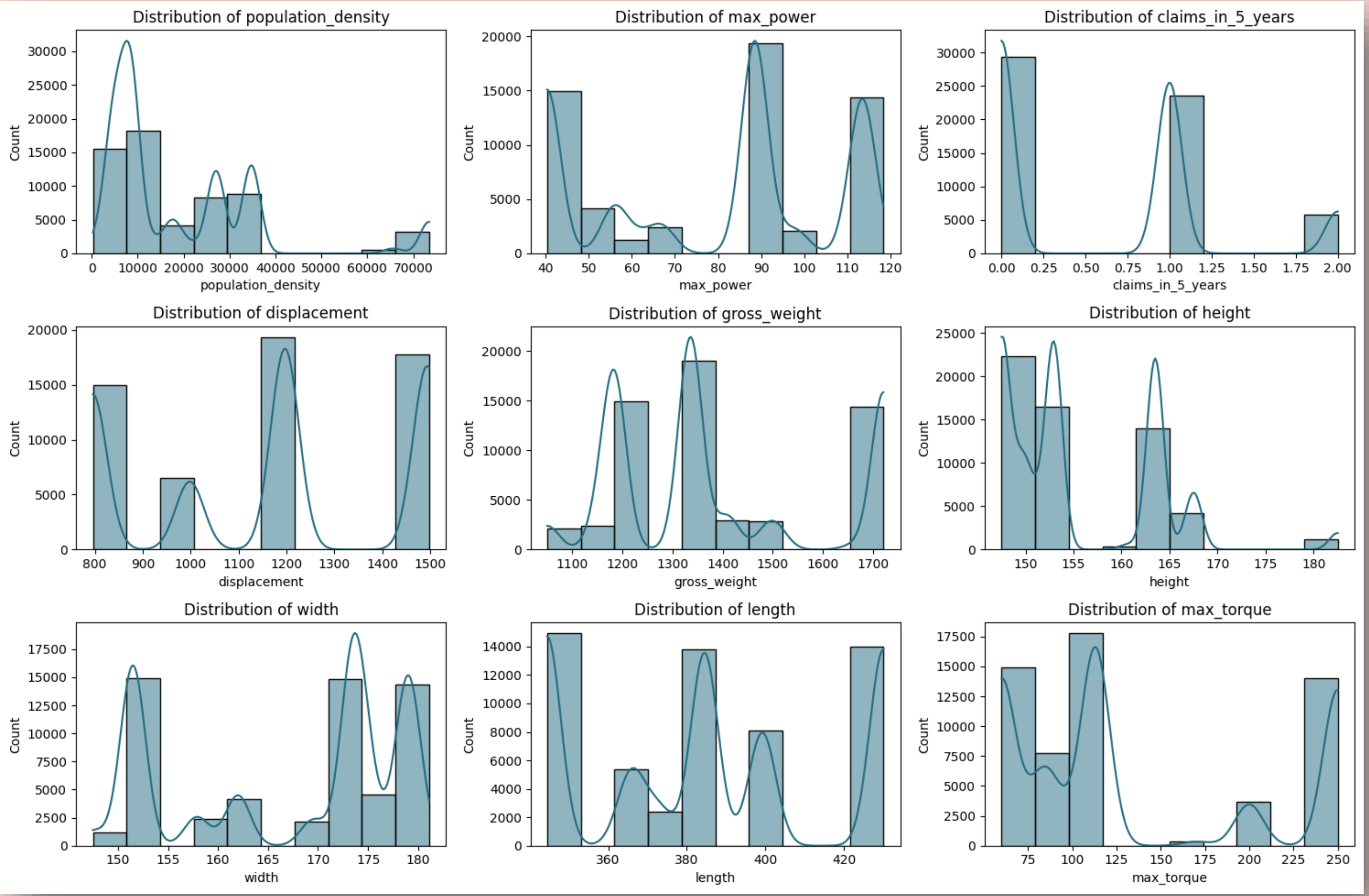
Heatmap of Correlation Matrix

- Identified **strong relationships** between variables like Customer Demographic, Displacement of Car, and claims.

- Also seen multi-collinearity among other features which can be seen in dark in heatmap, this shows that a simple Linear Regression Model is not a great choice for the data



Visualization of Data

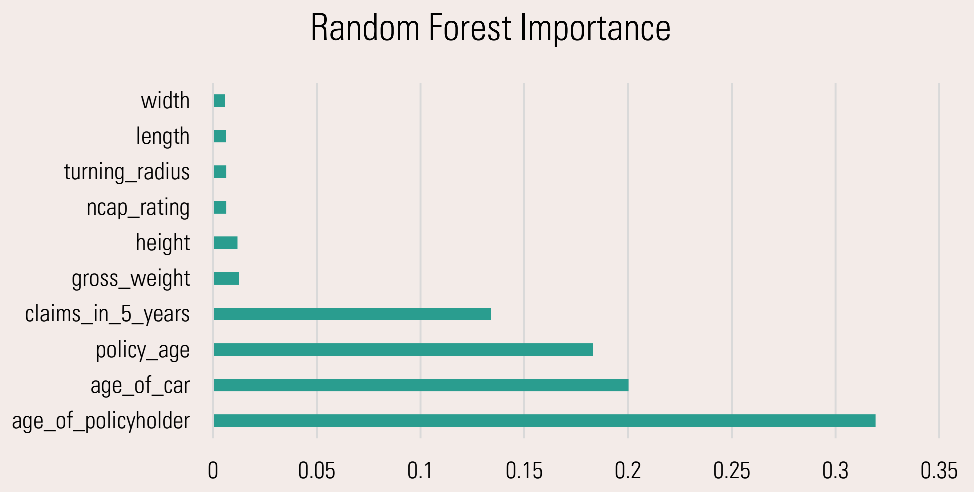


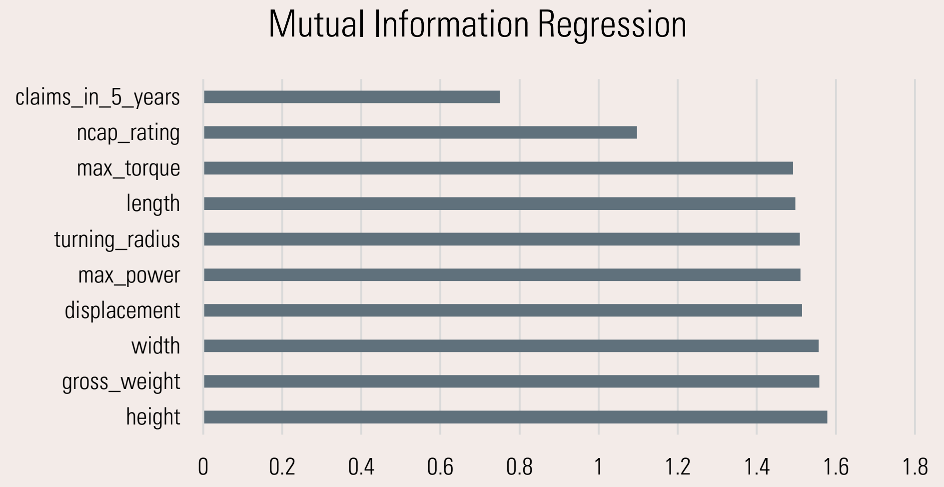
- ❑ Number of Claims in 5 years for each customer are one of 0,1,2 or 3, but considering only these are the only values possible, will create a bias in the model
- ❑ Most of the numerical features can be standardized
- ❑ We can see 3 categories in Length, vehicle is either small, mid or long

Transformations on Data

Variable	Transformation Applied	Reason	Effect
max_torque	Extract numeric values till "N", then standardization	The column contained "Nm" units, making it non-numeric	Enables numerical computation & brings it to a common scale for better model training
max_power	Extract numeric values till "b", then log transformation	The column contained "bhp" units, making it non-numeric & had a skewed distribution	Enables numerical computation, reduces skewness, and improves normality
length, width, height	Extract numeric values till "c", then standardization	The columns contained "cm" units, making them non-numeric	Enables numerical computation & prevents larger values from dominating the model
displacement	Extract numeric values till "c", then log transformation	The column contained "cc" units & had a right-skewed distribution	Converts it to a numerical format & stabilizes variance for better predictions
gross_weight	Standardization	The variable had a large range, requiring scaling	Prevents scale differences from affecting model training
is_esc, is_tpms, etc.	Convert "Yes"/"No" to 1/0	The column was categorical (binary) but should be numeric	Enables model to understand binary categorical variables
area, make, segment, etc.	One-Hot Encoding	The column was categorical with multiple levels	Prevents incorrect ordinal interpretation & allows better categorical representation

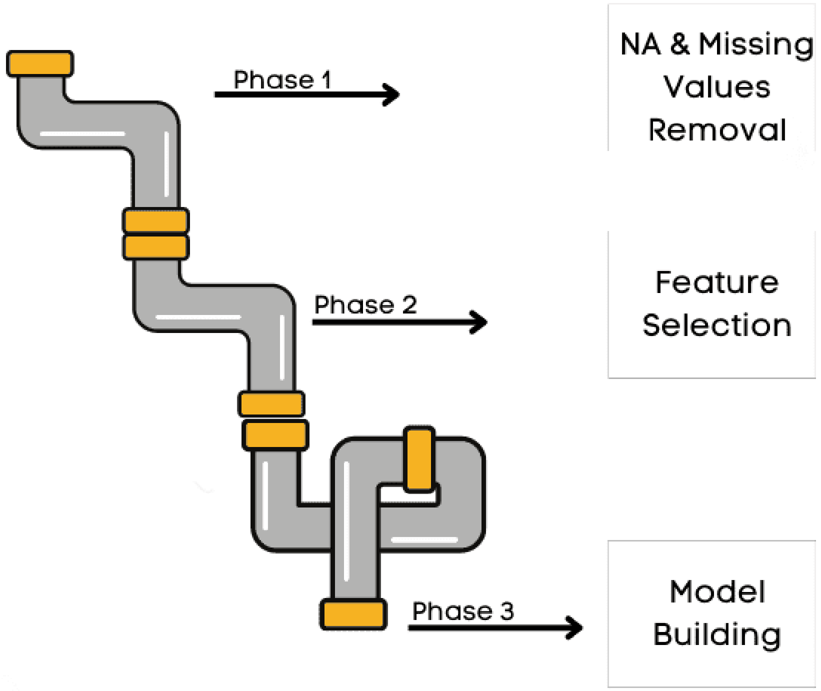
Important Features





1. Age of Policyholder
   - Younger drivers tend to have higher accident risks due to inexperience.
   - Older policyholders might have slower reaction times, affecting claim likelihood.
2. Age of Car
   - Older vehicles may have higher maintenance issues, leading to more claims.
   - Newer cars with advanced safety features might have fewer severe claims.
3. Claims in 5 Years
   - A history of frequent claims suggests a higher likelihood of future claims.
   - Past claim behavior is a strong predictor of risk profile.

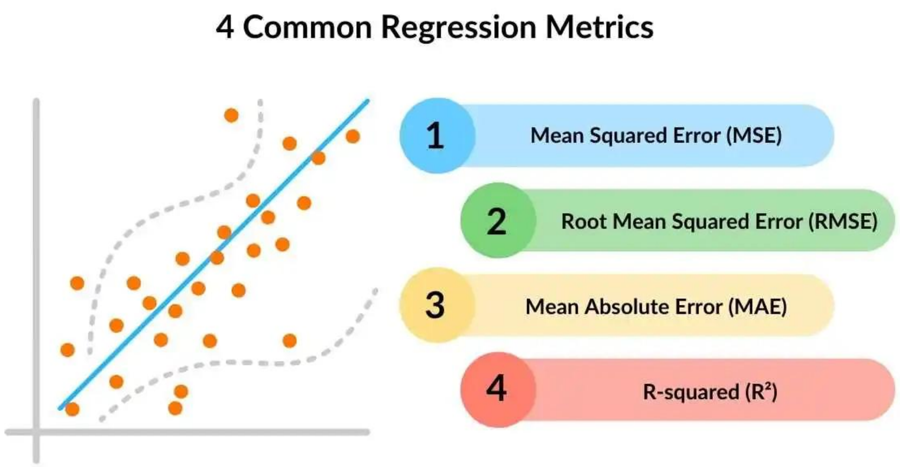
Model Building



Selected Machine Learning Models

- **Linear Regression** (Baseline Model)
- **Random Forest Regressor** (Captures non-linearity)
- **Gradient Boosting Regressor** (Boosted ensemble for improved accuracy)
- **Support Vector Regressor (SVR)** (Handles complex relationships)
- **XGBoost Regressor** (Optimized tree-based model)

Model Evaluation Metrics



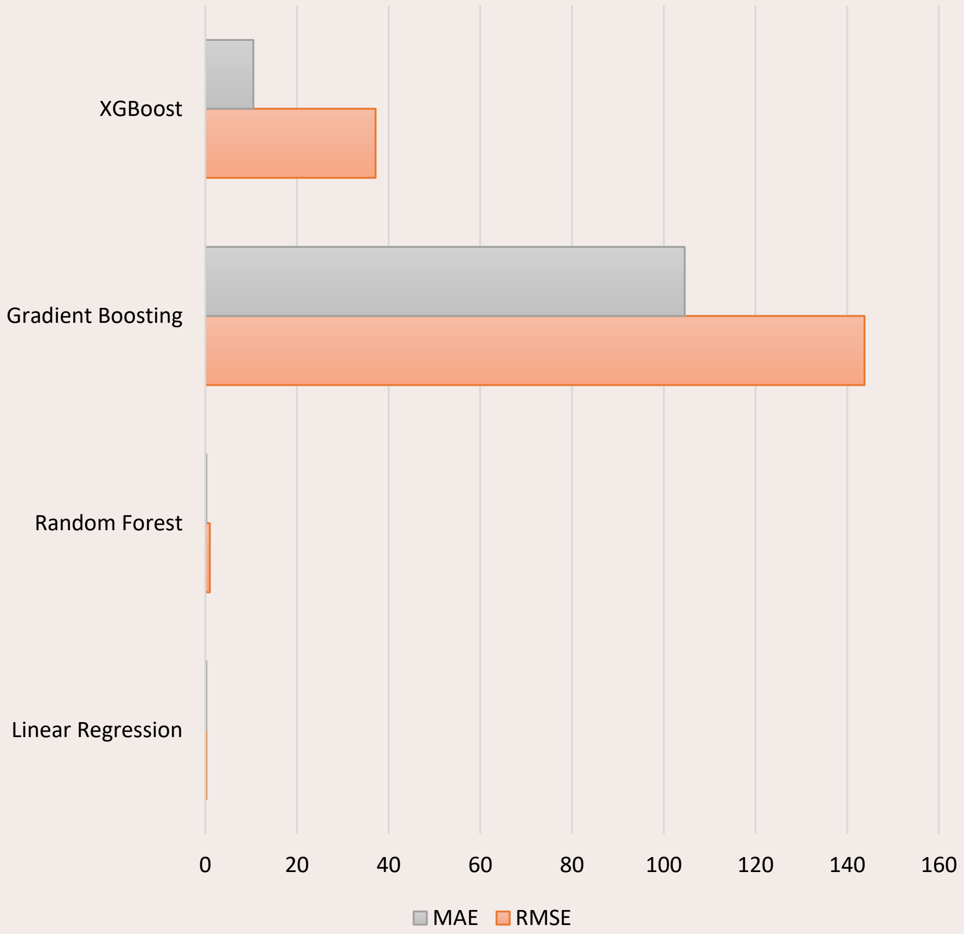
Metrics Used for Performance Evaluation:

- **Mean Squared Error (MSE)** – Measures average squared error.
- **Mean Absolute Error (MAE)** – Shows absolute deviation from actual claims
- **Root Mean Squared Error (RMSE)** – Penalizes large errors more than MSE
- **R-squared (R^2)** – Explains variance captured by the model

Model Evaluation Result

	MSE	RMSE	MAE	R^2 Score
Linear Regression	0.069293	0.263235	0.207078	1
Random Forest	1.031695	1.015724	0.25302	1
Gradient Boosting	20682.22	143.8131	104.5614	0.993324
Support Vector Regressor	3929821	1982.378	1224.092	-0.26846
XGBoost	1379.906	37.14708	10.5019	0.999555

Visualization of Results

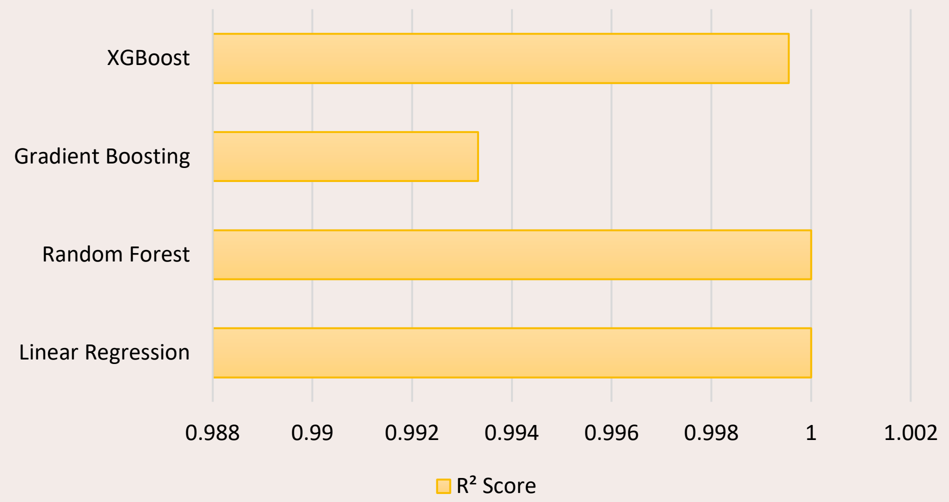


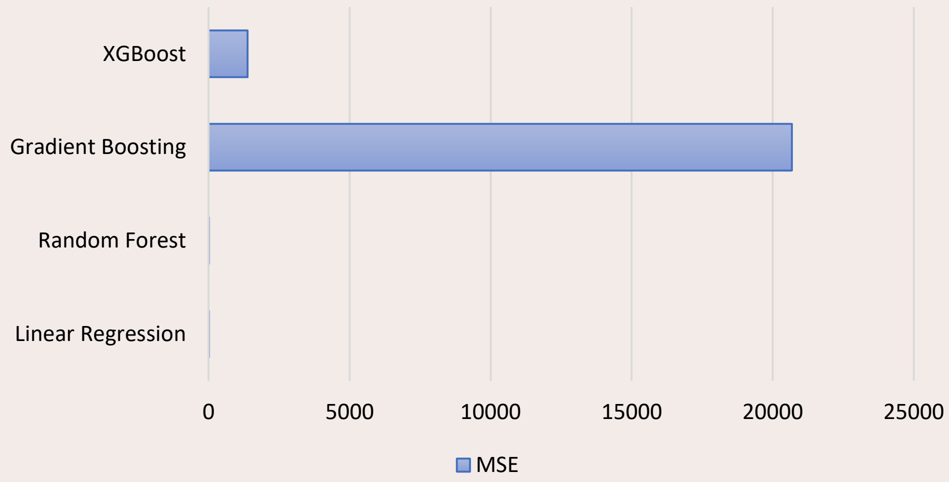
We evaluate models based on **MSE, RMSE, MAE, and R^2 Score**:

- **Linear Regression**: Lowest MSE and RMSE (0.07, 0.26), R^2 = **1.00**
- **Random Forest**: Higher MSE and RMSE (1.03, 1.02), R^2 = **1.00**
- **Gradient Boosting**: Very high MSE (20,682), but **R^2 = 0.993**
- **Support Vector Regressor (SVR)**: Poor performance, **negative R^2 (-0.268)** → Not suitable
- XGBoost: Very strong performance (MSE = 1379, RMSE = 37.14, MAE = 10.50, R^2 = 0.9996)

**SVR Removed from chart because of very high RMSE, MSE and Negative R^2

Visualization of Results





We evaluate models based on **MSE, RMSE, MAE, and R^2 Score**:

- ❑ **Linear Regression**: Lowest MSE and RMSE (0.07, 0.26), R^2 = **1.00**
- ❑ **Random Forest**: Higher MSE and RMSE (1.03, 1.02), R^2 = **1.00**
- ❑ **Gradient Boosting**: Very high MSE (20,682), but **R^2 = 0.993**
- ❑ **Support Vector Regressor (SVR)**: Poor performance, **negative R^2 (-0.268)** → Not suitable
- ❑ XGBoost: Very strong performance (MSE = 1379, RMSE = 37.14, MAE = 10.50, R^2 = 0.9996)

**SVR Removed from chart because of very high RMSE, MSE and Negative R^2

Why Choose XGBoost?

1. High Accuracy & Low Error
    - **R^2 Score: 0.9996**, indicating a near-perfect fit while avoiding complete overfitting.
    - **MSE: 1379, RMSE: 37.14, MAE: 10.50**, showing significantly lower prediction errors compared to Gradient Boosting and SVR.
2. Better Generalization
    - While **Linear Regression** has an **R^2 of 1.00**, its extremely low errors (MSE = 0.07) suggest **overfitting or data leakage**.
    - XGBoost balances accuracy while **maintaining robust generalization**.
3. Handles Complex Data Relationships
    - XGBoost **captures non-linear patterns** better than Linear Regression and SVR.
    - Unlike **Random Forest**, it reduces variance and avoids overfitting through boosting techniques.
4. Efficient & Scalable
    - XGBoost is optimized for **speed and efficiency**, making it suitable for **large datasets**.
    - It uses **regularization** (L1/L2) to prevent overfitting.

Key Insights from the Model

Policyholder and Vehicle Factors Matter

- **Older policyholders & vehicles** → Higher claim severity
- **Frequent past claims** → Strong indicator of future claim severity
- **Heavy and high-powered vehicles** → More expensive claims due to repair costs

Top Features Driving Claim Severity

- **Age of Policyholder & Age of Car:** Older individuals and vehicles correlate with higher claims.
- **Claims in Last 5 Years:** Strong predictor of future claims, highlighting the importance of claim history in risk assessment.
- **Vehicle Weight & Dimensions:** Heavier, larger vehicles incur **higher repair costs**.
- **NCAP Safety Rating:** Lower safety ratings link to **higher claim amounts** due to severe damage/injuries.

How This Supports Pricing & Risk Management

Refined Premium Pricing

- Identify **policyholders prone to high claims** before issuing policies.
- Adjust coverage plans for **high-risk vehicles**.

Better Underwriting Decisions

- Identify **policyholders prone to high claims** before issuing policies.
- Adjust coverage plans for **high-risk vehicles**.

Risk Mitigation Strategies

- Offer **incentives for safer vehicles** (higher NCAP rating).
- **Encourage responsible driving behavior** through dynamic pricing (e.g., usage-based insurance).

Limitations & Future Improvements

Missing Key Behavioral Data

- Driving habits (mileage, speeding, accident history) not included.
- Future improvement: **Integrate telematics data**.

External Factors Not Considered

- Weather, road conditions, economic factors could refine predictions.
- Solution: **Incorporating external datasets for a holistic approach**.

Explore Advanced Models

- **Generalized Linear Models (GLM)** or deep learning can be used to enhance prediction accuracy.

Thank you

Anant Kumar Verma
Christ University
+91 7453 015 236
anant.kumar@arts.christuniversity.in