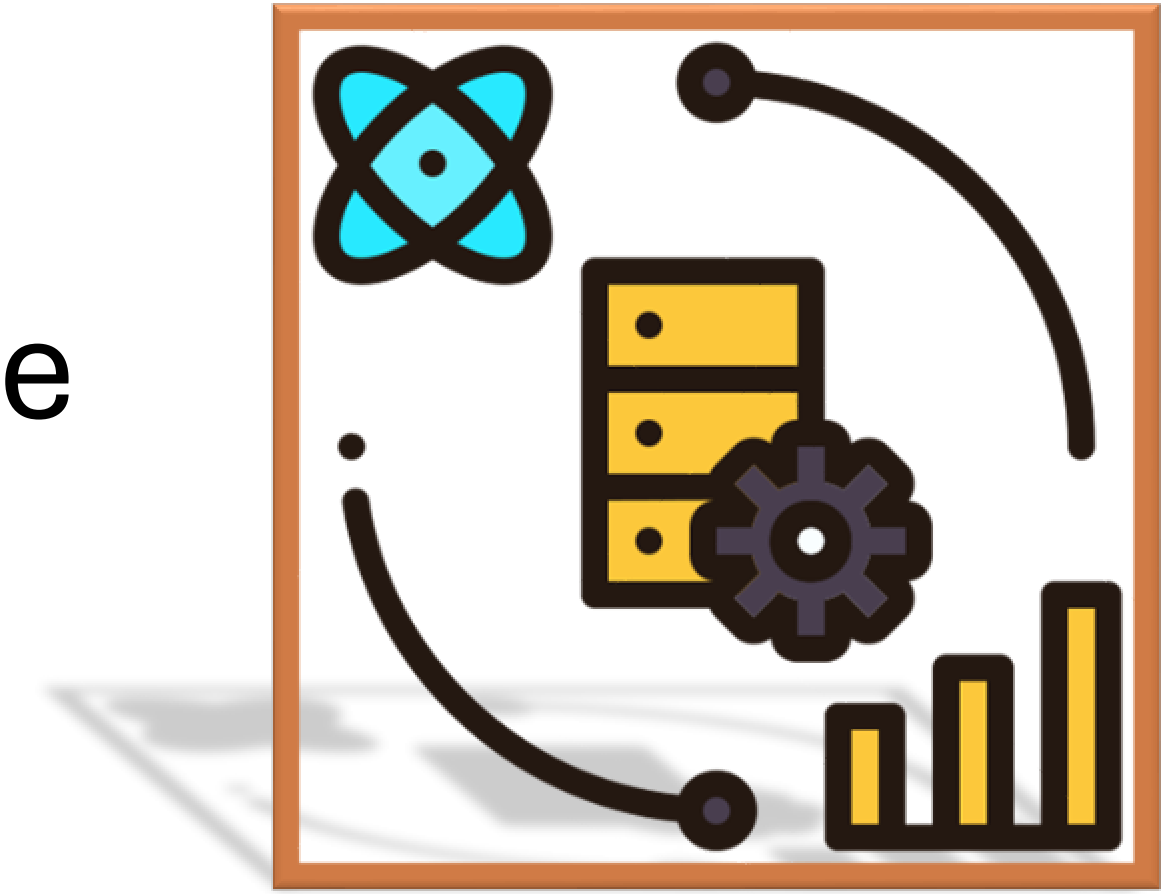
Predicting Insurance Policy Renewal

Garrett Sparks, Alper Tepebas, Yichong Wu



Car Insurance Renewal Prediction

- Goal: Predict how likely a customer is to renew their car insurance policy
- Focused on comparing different machine learning models to find the most accurate one
- Models Evaluated:
  - Linear Regression
  - Random Forest
  - Gradient Boosting

Data Source

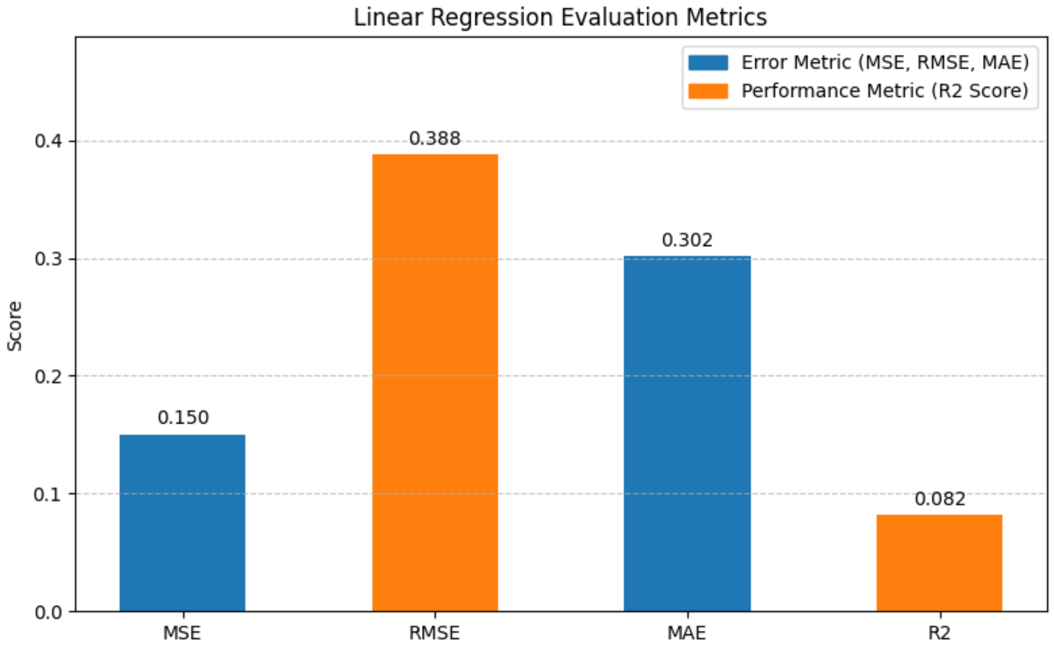
- *Dataset of an actual motor vehicle insurance portfolio* from Mendeley Data
- This data has **105,555 rows of individual policy transactions**
- **30 different variable columns**
- The data comes from research performed by *Universitat de Valencia* (University of Valencia)
- We leveraged our data 80%/20% training vs test data

Data Preprocessing

- Our data has some Qualitative Values, we use binary numbers for those values
- Setting a Reference Date
- Feature Engineering
  - Created a feature to show a driver's experience based on given info such as the date a driver got their license and comparing it to reference date
- Filtering Outliers
- Normalization
- Dropping Columns
  - Example is ID number, as it has no real value to if or if not a customer renews their policy
  - We would keep a column like a customer's age
- Followed the Consecutive Rule

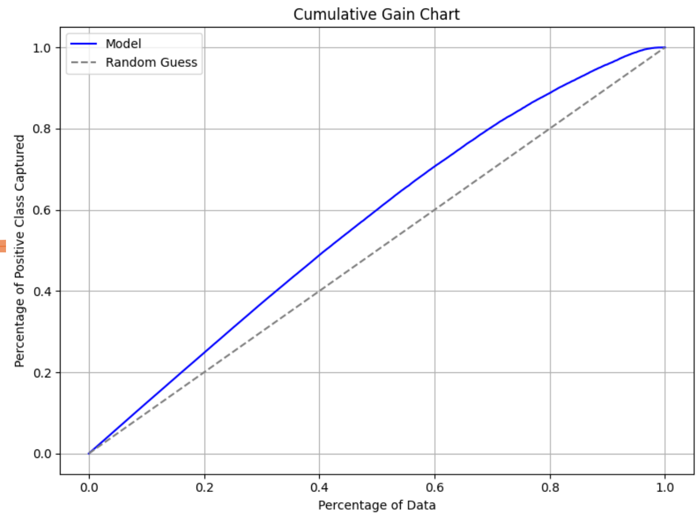
Linear Regression

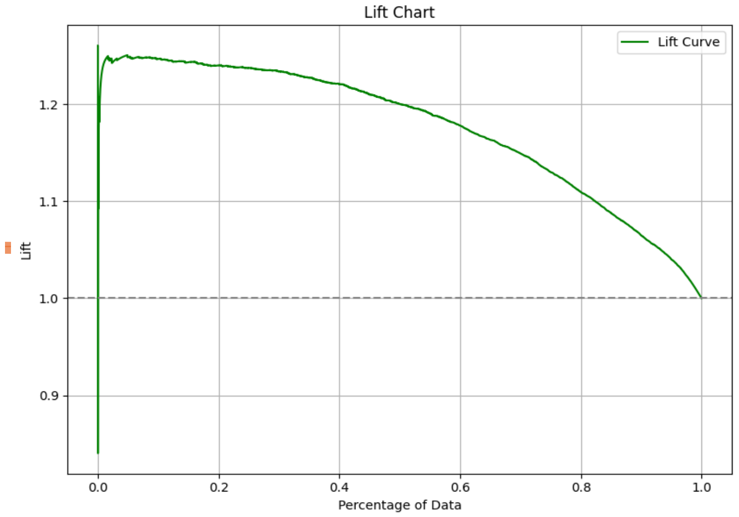
- **Extremely low R^2 score (0.082):**
  - The model only explains about 8% of what's happening in the data
- **MSE (0.150):**
  - The model has a target of [0,1], making the MSE acceptable, not good nor bad
- **RMSE (0.388):**
  - The model's predictions are almost 40% off from the true values
  - A lot of guessing instead of accurate prediction
- **MAE (0.302):**
  - Every prediction is wrong by about 30% on average
  - Errors happen consistently across many examples
- **Linear regression is a bad fit for this dataset**

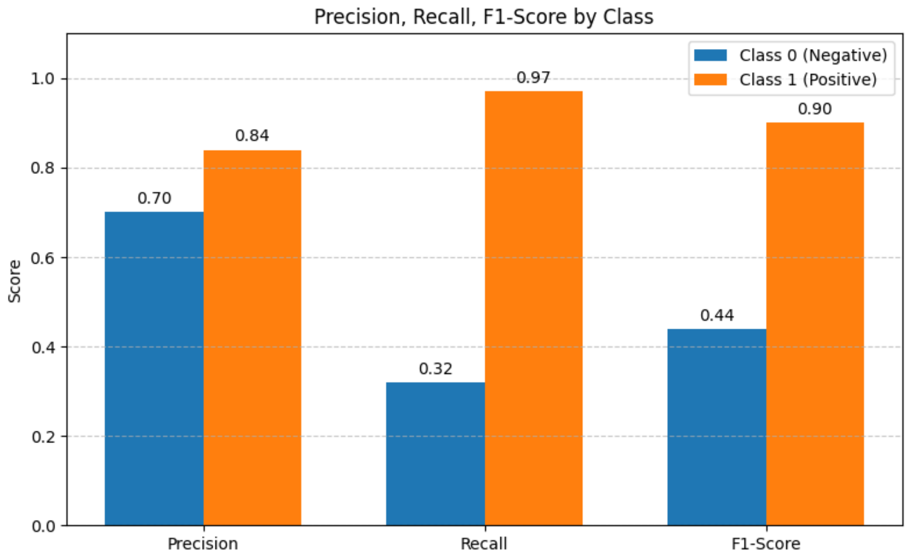


Random Forest

- **Approach**:
  - 100 trees (robust ensemble), Class balancing (class_weight='balanced')
- **Performance Metric**:
  - Accuracy: 0.8312
- The model is strong at identifying positive cases.
- However, it struggles significantly with negative cases.
- Random Forest model shows underfitting toward churners, with high bias and poor recall for customers who leave, while performing well for renewals.

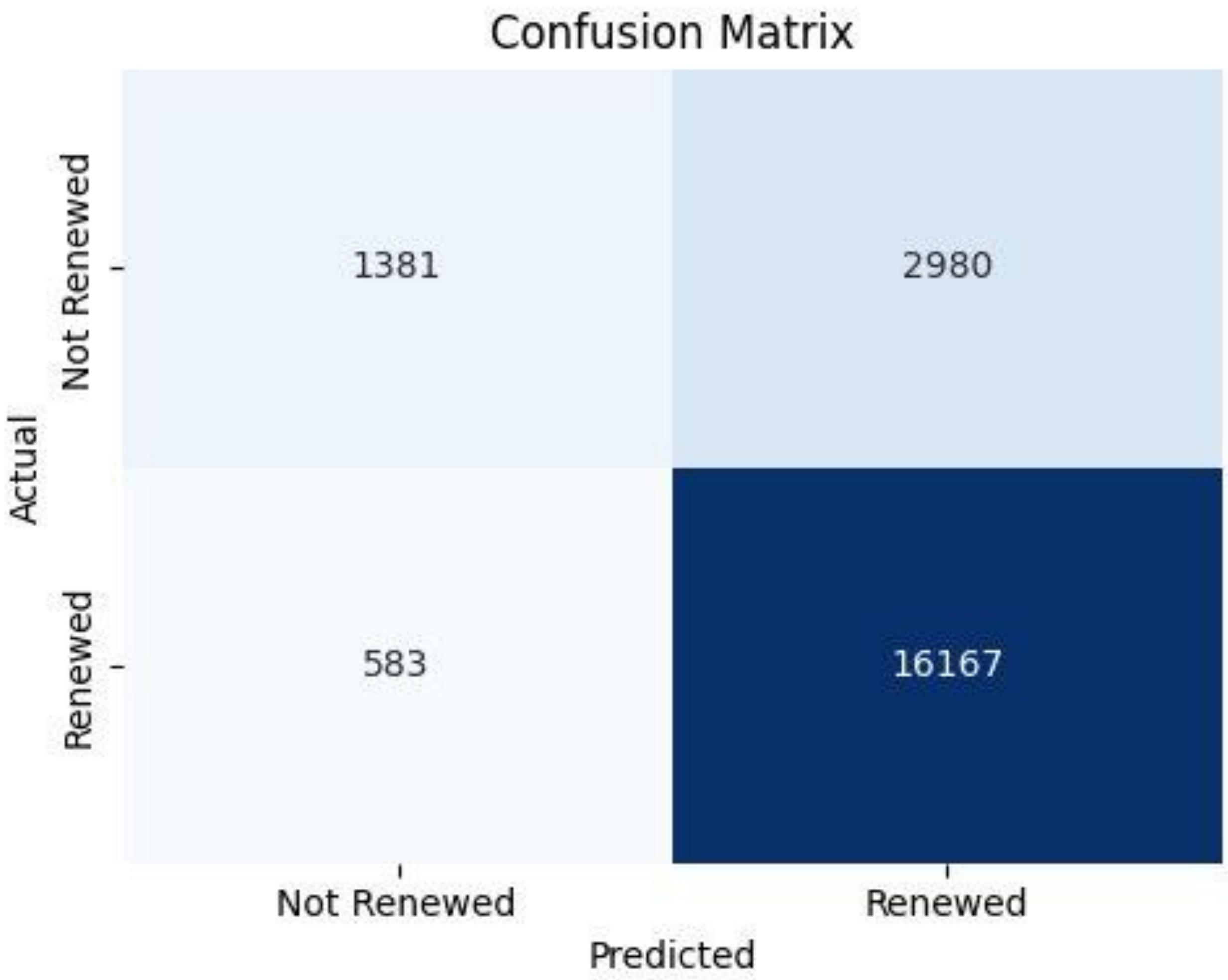






Random Forest

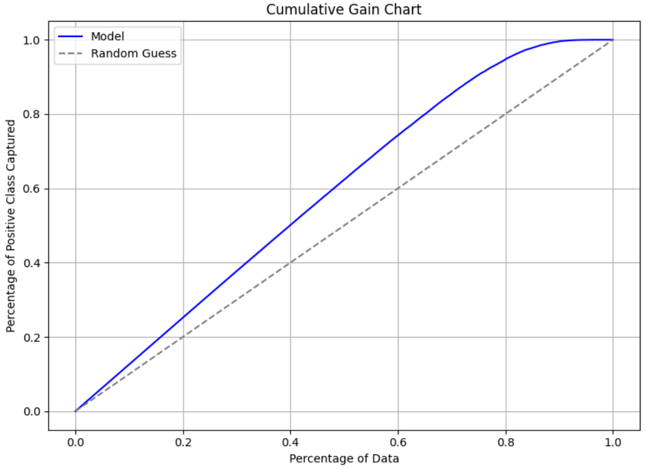
Confusion Matrix on Test Data

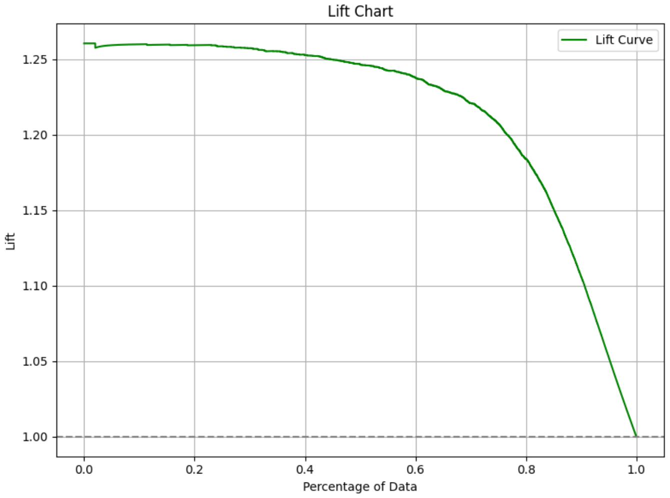


Confusion Matrix

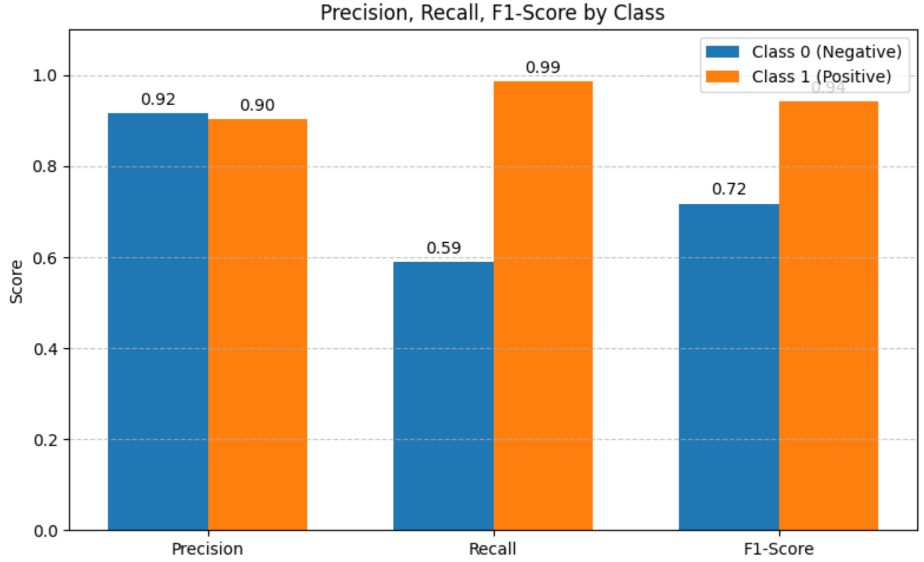
Actual \ Predicted	Not Renewed	Renewed
Not Renewed	1381	2980
Renewed	583	16167

Gradient Boosting



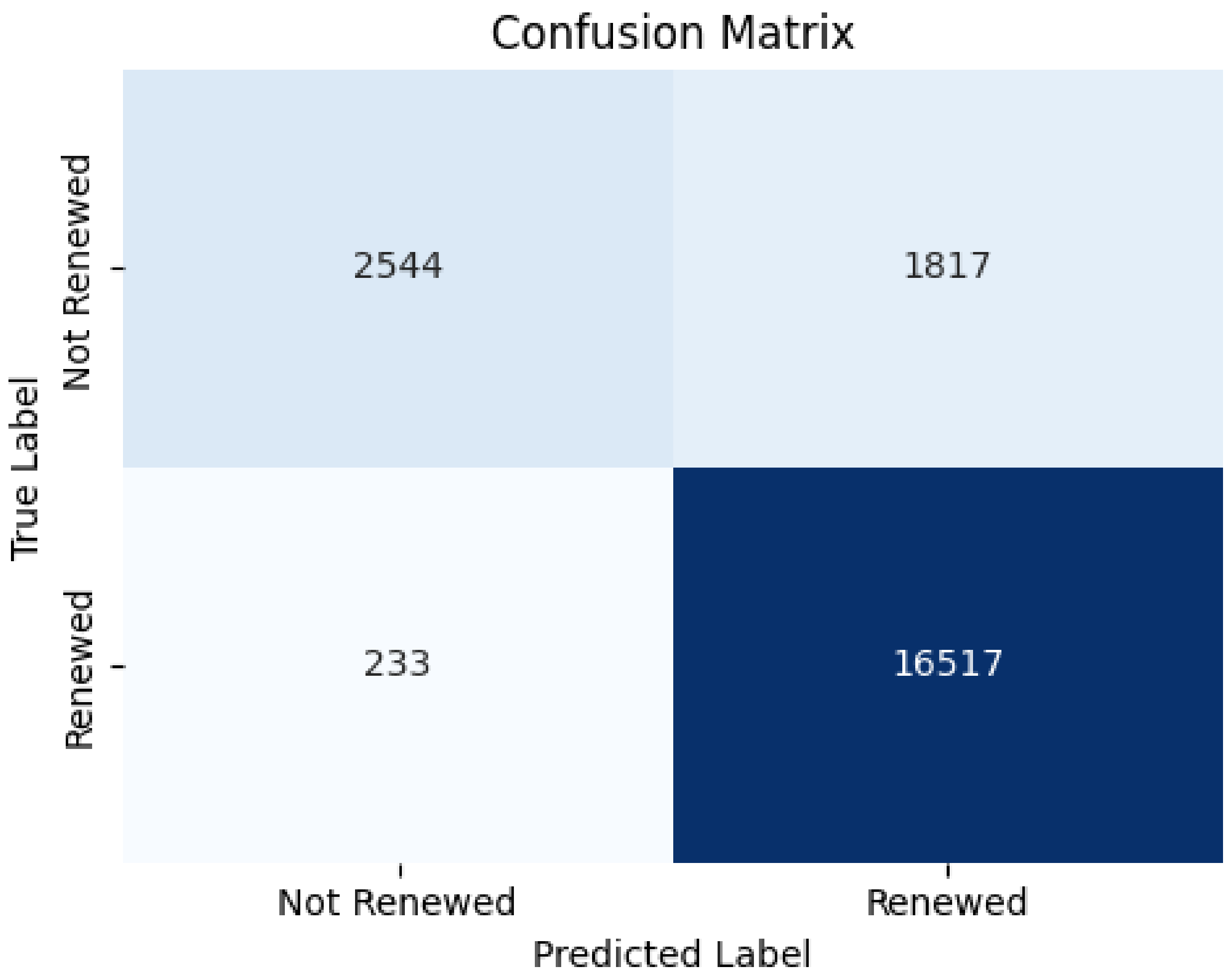


- **Approach**:
  - 591 trees, learning rate 0.036, max depth 9; min_samples_split=180, min_samples_leaf=30, subsample≈0.92
- **Performance Metric**:
  - Accuracy: 0.9029
- The model does very well overall, with high accuracy and strong scores in almost all areas.
- Model fits well without signs of major underfitting or overfitting.
- It improved detection of customers who would leave, compared to other models.



Gradient Boosting

Confusion Matrix on Test Data



Confusion Matrix

True Label \ Predicted Label	Not Renewed	Renewed
Not Renewed	2544	1817
Renewed	233	16517

Hyperparameter Tuning

- After tuning, the accuracy increased to over 90%, compared to around 87% for the original Gradient Boosting.
- We used randomized search to find the following parameters:
  - number of trees
  - Learning rate
  - Max depth
  - Minimum number of samples per leaf
  - Minimum number of samples split
  - Samples used for fitting each tree
  - Random state

Comparison

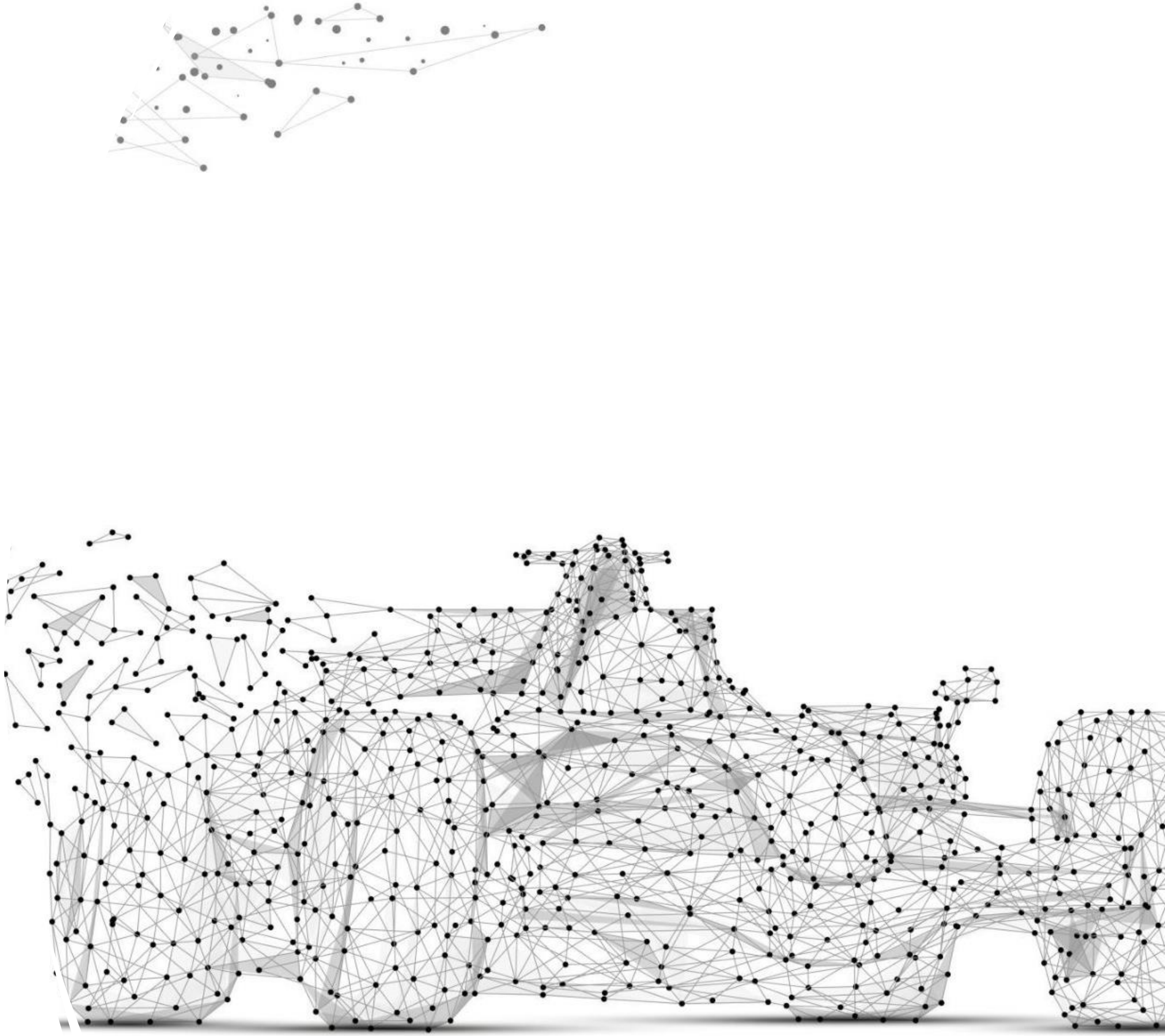
Hyperparameter-tuned Gradient Boosting outperformed all other models.

Tuned model gave us the highest accuracy out of all the models and made a big impact.

Non-tuned Gradient Boosting also performed better than Random Forest, but not as strongly as the hyperparameter-tuned version.

Thank You

- Garrett Sparks
- Alper Tepebas
- Yichong Wu



Works Cited

- Bisht, Ankit. “ML | Classification vs Regression - GeeksforGeeks.” *GeeksforGeeks*, 8 Jan. 2019, www.geeksforgeeks.org/ml-classification-vs-regression/.
- GeeksForGeeks. “Hyperparameter Tuning.” *GeeksforGeeks*, 7 Dec. 2023, www.geeksforgeeks.org/hyperparameter-tuning/.
- Lledó, Josep (2023), “Dataset of an actual motor vehicle insurance portfolio”, Mendeley Data, V1, doi: 10.17632/5cxyb5fp4f.1
- ---. “ML - Gradient Boosting.” *GeeksforGeeks*, 25 Aug. 2020, www.geeksforgeeks.org/ml-gradient-boosting/.
- GeeksforGeeks. “Random Forest Algorithm in Machine Learning.” *GeeksforGeeks*, 12 July 2024, www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/.
- Segura-Gisbert, J., Lledó, J. & Pavía, J.M. Dataset of an actual motor vehicle insurance portfolio. *Eur. Actuar. J.* 15, 241–253 (2025). https://doi.org/10.1007/s13385-024-00398-0
- ---. “Top 65+ Data Science Projects with Source Code.” *GeeksforGeeks*, 8 Feb. 2024, www.geeksforgeeks.org/top-data-science-projects/. Accessed 29 Apr. 2025.
- Gupta, Mohit. “ML | Linear Regression - GeeksforGeeks.” *GeeksforGeeks*, 13 Sept. 2018, www.geeksforgeeks.org/ml-linear-regression/.
- Kumar, Ajitesh. “Mean Squared Error or R-Squared - Which One to Use?” *Data Analytics*, 30 Sept. 2020, vitalflux.com/mean-square-error-r-squared-which-one-to-use/.