Connor Ray & Jacob Stewart
Group 7

Predicting auto claim severity and classifying clients using XGBoost

Contents

Motivation

Claims are both the reason the insurance exist and the enemy of its profitability.

In order for an insurer to remain profitable, it must have a proper understanding of the risks it accepts and price it appropriately.

Our project set out to be a tool for insurers to use to gain data-driven insight on client risk-level and claim severity.

Research Questions

	Question	Approach
1	Using the data, how closely can we predict an individual's total claim severity?	Identify a target variable to use as a predictor of claim severity.
2	What factors most strongly influence claim severity?	Evaluate what features contribute the most when predicting claim severity.
3	Based on our data, can we assign individuals into groups based on their risk level?	Use the severity distribution in our dataset to classify individuals.

Dataset

Auto Insurance Claims
Source: Kaggle
Size: 9,134 auto insurance claims with 34 attributes
Adjustment: Originally collected in 2011; all dollar amounts updated to 2024 values

What the Dataset Contains:

Demographic Variables
- Age, Gender, Marital Status, Education, Income, State

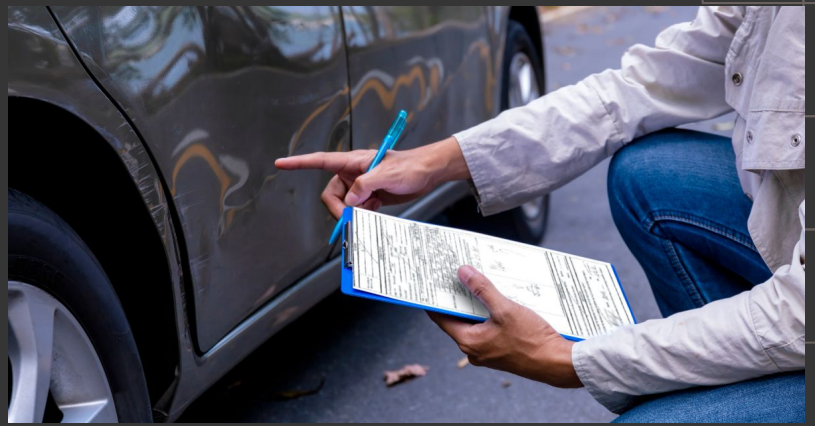
Policy Information
- Policy Type, Coverage Level, Sales Channel, Months Since Last Claim

Vehicle & Location Features
- Vehicle Class, Vehicle Size, Location Index

Claim-Related Fields
- Incident Type, Incident Severity, Total Claim Amount (target variable)

Why This Dataset Matters:

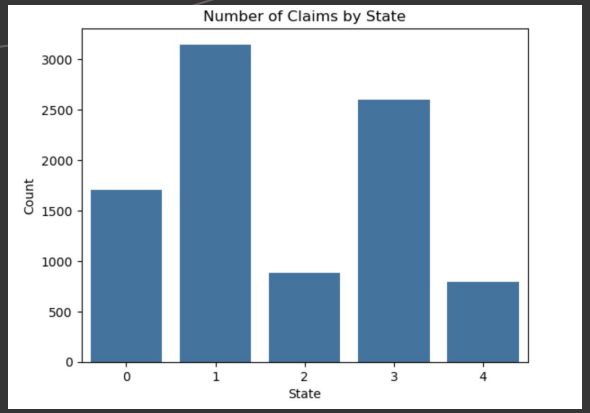
This dataset reflects the real factors insurers use to understand and manage claim severity. By analyzing patterns across demographics, policy details, and vehicle characteristics, insurers can price policies more accurately, identify high-risk customers, and improve decisions around underwriting and reserving.

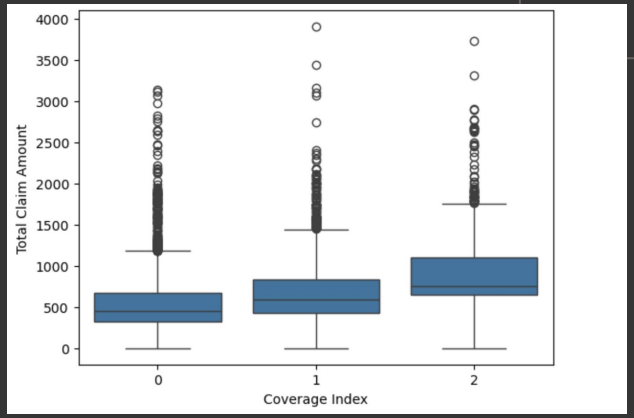


Exploratory Data Analysis

Since the dataset is primarily categorical, our EDA is qualitative/frequency based.

We relied on categorical distribution and relationship to the target variable to make decisions about what features to include in our model.

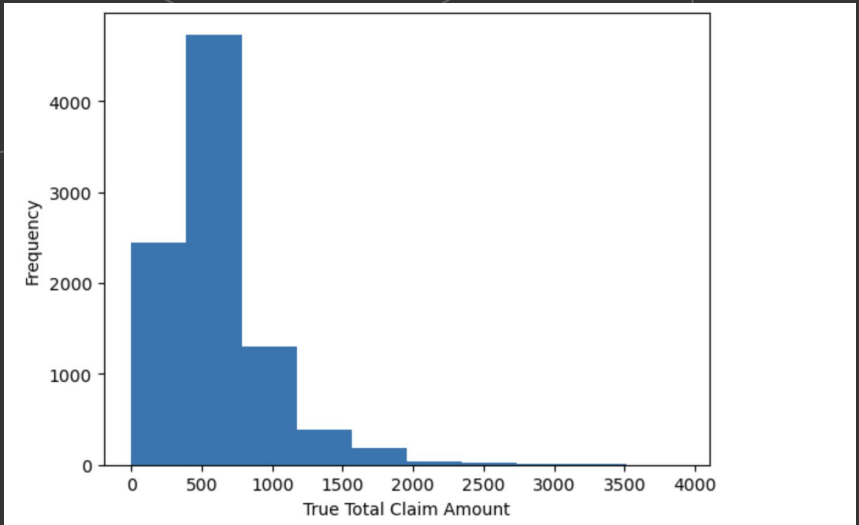





```
Education Index
2    0.300854
1    0.293519
0    0.287059
3    0.081125
4    0.037443
Name: proportion, dtype: float64
```

Target Variable: Total Claim Amount

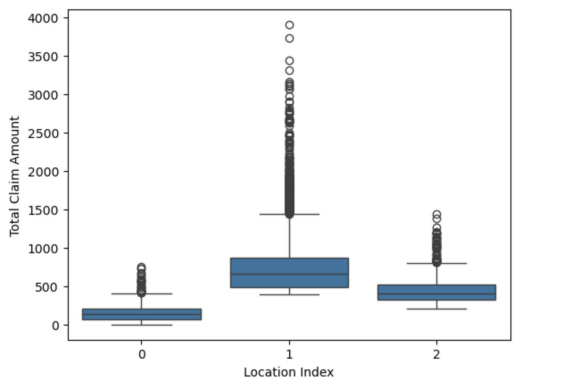
```
Max Total Claim Amount 3905.87
Min Total Claim Amount 0.13
Mean Total Claim Amount 586.0199135099627
Variance Total Claim Amount 153801.34252025004
Standard Deviation Total Claim Amount 392.1751426598198
```

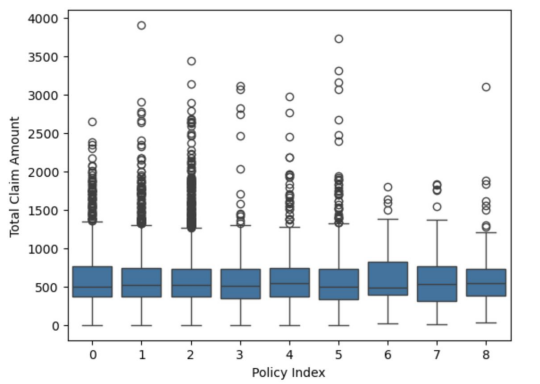


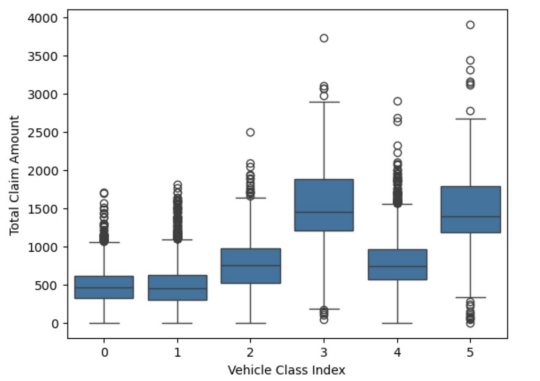
Non-normal distribution:
XGBoost supports non-normally distributed target variables because tree models do not assume any distribution

Exploratory Data Analysis

Features that held similar distribution across groups will not be strong predictors and will rather add noise to the model, so we removed features with similar distributions across groups.







Cardinality & Dominant Groups

```
Marital Status Index
1    0.580031
0    0.270090
2    0.149880
Name: proportion, dtype: float64
```

```
Vehicle Class Index
1    0.505912
0    0.206481
4    0.196628
2    0.052989
5    0.020145
3    0.017845
Name: proportion, dtype: float64
```

State	5
Coverage	3
Coverage Index	3
Education	5
Education Index	5
Employment Status	5
Employment Status Index	5
Gender	2
Income	5694
Location	3
Location Index	3
Marital Status	3
Marital Status Index	3
Monthly Premium Auto	202
Months Since Last Claim	36
Months Since Policy Inception	100
Number of Open Complaints	6
Number of Policies	9
Policy Type	3
Policy Type Index	3
Policy	9
Policy Index	9
Renew Offer Type	4
Total Claim Amount	4989
Vehicle Class	6
Vehicle Class Index	6
Vehicle Size	3
Vehicle Size Index	3
Income Index	5
State Index	5
Gender Index	2

dtype: int64

Model Training, Parameter Tuning, & Results

Model Training

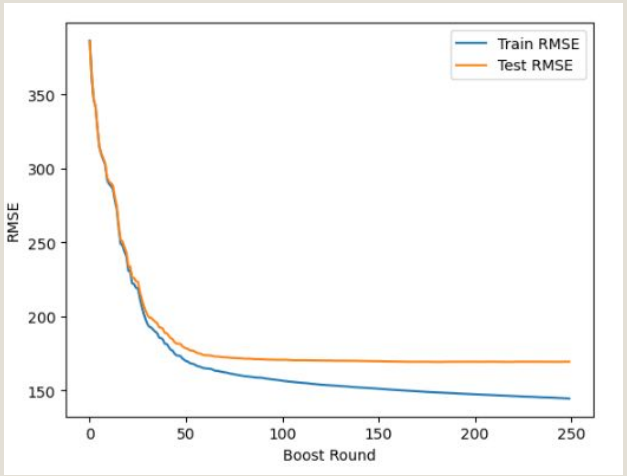
- Model used: **XGBoost Regressor** (tree-based, handles nonlinear patterns well)
- Train/test split: **80% train, 20% test**
- Used only **indexed / encoded features** to simplify structure
- Removed weak predictors to reduce noise and avoid overfitting

Parameter Tuning

- **Initial parameters:** max_depth = 5, eta = 0.1, subsample = 0.5, colsample_bytree = 0.5
- Lowering Max_Depth to 4 improved generalization
- Removing **Education Index** reduced redundancy with Income
- **5-fold cross-validation** identified **~50 boosting rounds** as optimal

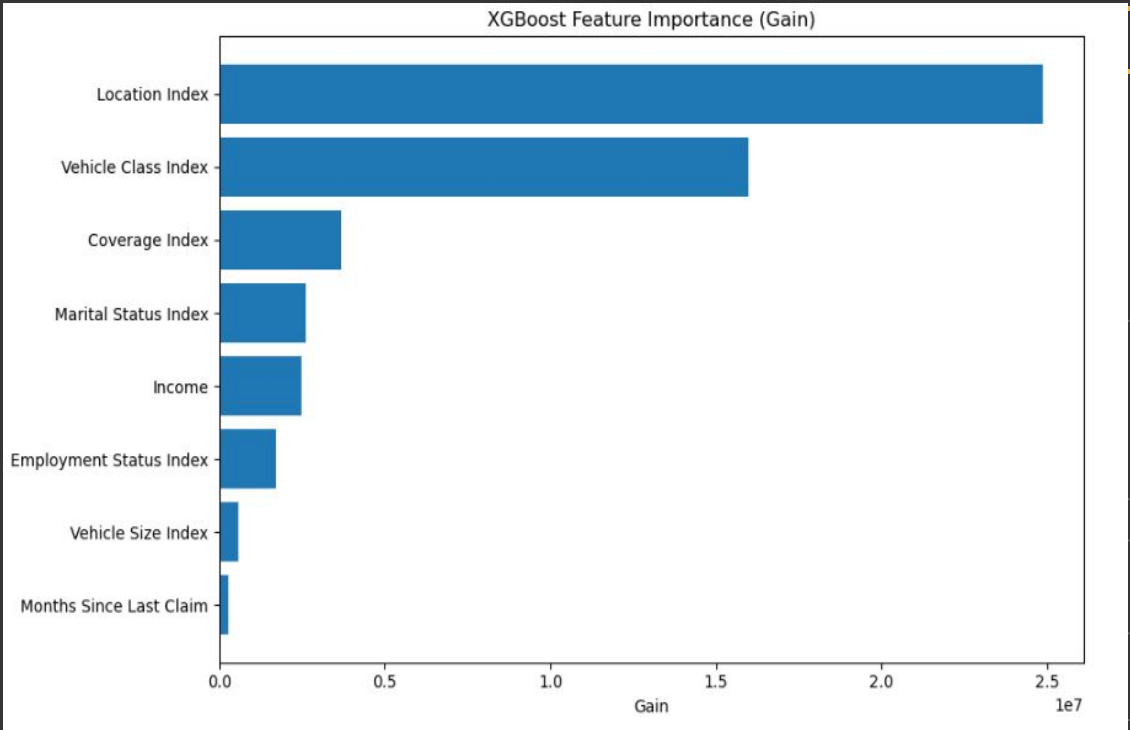
Final Model Performance

```
RMSE: 161.66092195604875
R2: 0.8257660519513981
```



	Coverage Index	Employment Status Index	Income	Location Index	Marital Status Index	Months Since Last Claim	Vehicle Class Index	Vehicle Size Index	Actual Total Claim Amount	Predicted Total Claim Amount
2558	0	0	0	1	0	26	1	2	440.64	664.603394
7882	0	3	20214	1	2	3	4	1	1117.80	864.866455
608	0	1	125834	0	1	47	0	2	100.61	99.488953
1242	0	0	0	1	1	16	4	1	771.12	882.568420
652	0	0	0	1	0	39	1	1	602.64	666.520020
...	...	...	...	...	...	...	...	...	...	...
7838	0	1	106134	2	2	9	0	1	304.53	331.034851
6483	0	1	85856	0	1	14	4	1	248.37	202.122055
2893	0	1	71117	0	1	28	1	1	159.30	127.656570
2531	0	0	0	1	0	26	1	2	440.64	664.603394
4903	0	1	38553	1	1	19	4	2	1076.70	789.784851

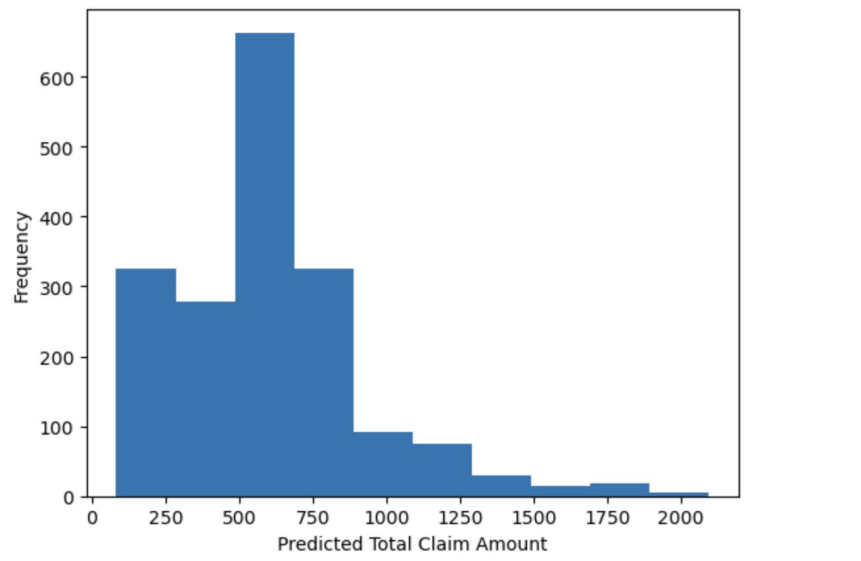
Feature Importance



Gain is the average improvement in loss across all splits where the specific feature is used.

A high gain shows that the feature contributed a large reduction in the loss function across all the times where it was the criteria the data was split on.

Individual Risk-Level Classification



Using these percentiles as bounds, we labeled each observation into a relative risk level based on what bin its total claim amount fell into.

This method allowed us to partition the dataset into groups such as Low Risk, Medium Risk, High Risk, and Very High Risk based solely on the actual loss amounts observed in the data.

```
percentiles = np.percentile(Y_predictions, [1, 25, 50, 75, 90, 95, 99])
percentiles
```

```
array([ 103.98663818,  348.92251587,  541.81512451,  773.83859253,
        966.65595703, 1177.73405762, 1767.03005371])
```

	Coverage Index	Employment Status Index	Income	Location Index	Marital Status Index	Months Since Last Claim	Vehicle Class Index	Vehicle Size Index	Actual Total Claim Amount	Predicted Total Claim Amount	Predicted Customer Risk Level
2558	0	0	0	1	0	26	1	2	440.64	664.603394	High
7882	0	3	20214	1	2	3	4	1	1117.80	864.866455	Extreme
608	0	1	125834	0	1	47	0	2	100.61	99.488953	Low
1242	0	0	0	1	1	16	4	1	771.12	882.568420	Extreme
652	0	0	0	1	0	39	1	1	602.64	666.520020	High
...	...	...	...	...	...	...	...	...	...	...	...
7838	0	1	106134	2	2	9	0	1	304.53	331.034851	Low
6483	0	1	85856	0	1	14	4	1	248.37	202.122055	Low
2893	0	1	71117	0	1	28	1	1	159.30	127.656570	Low
2531	0	0	0	1	0	26	1	2	440.64	664.603394	High
4903	0	1	38553	1	1	19	4	2	1076.70	789.784851	Extreme

Conclusion

1. How closely can we predict an Individual's total claim severity?
   - **XGBoost** predicted claim severity accurately, with **RMSE ≈ 160** and **R^2 ≈ 0.83**.
   - **Model performance** shows severity can be reliably estimated relative to the dataset's average claim amount.
2. What factors most strongly influence claim severity?
   - **Top Predictors**: Location, Vehicle Class, Coverage Level, and Income showed the highest gain and contributed most to reducing model error.
   - **Low-Value Predictors**: Gender, State Index, and several policy identifiers provided little predictive signal and mostly added noise.
3. Based on our data, can we assign individuals into groups based on their risk level?
   - Used **quartiles** of Total Claim Amount to create **Low**, **Medium**, **High**, and **Very High Risk** groups.
   - These bins effectively separate customers by **observed claim severity** and support **practical risk classification.**

This analysis gives us a clearer understanding of what actually drives severe claims, allowing us to make more informed decisions about pricing and underwriting.

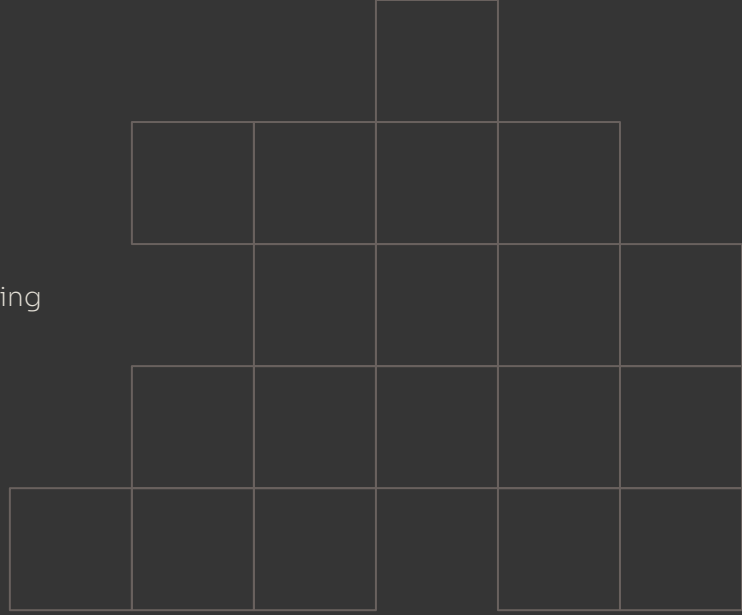
Things We Can Improve On

Limited Dataset:

- Our dataset did not include common features like Age, Vehicle Specifics, Driving Record, etc.
- Our dataset only included claims. A dataset with observations without a claim would help predict whether or not a claim is expected.
- Additional charges: towing, rental cars, etc.

Feature Transformation:

- Normalizing our target feature could improve prediction as XGBoost can sometimes favor outliers in predictions.



Thank you