Leveraging Historical Crash Data for Enhanced Insurance Pricing

DataMaven

Our Team

Pacifique Iradukunda

Risk Modeling Expert

Wenchun Zhang

Principal Investigator

Elie Niringiyimana

Field Specialist

Jenevieve Zhang

Deployment Strategist

Sunday Nwanyim

Machine Learning Engineer

Introduction-CRISP

Business Understanding

Data Science: A Business Value Proposition

Addressing Business Problems with Analytical Solutions

Leveraging Data for Insurance Excellence

- Enhancing Policy Pricing, Resource Allocation, and Accident Prevention

Historical Reports

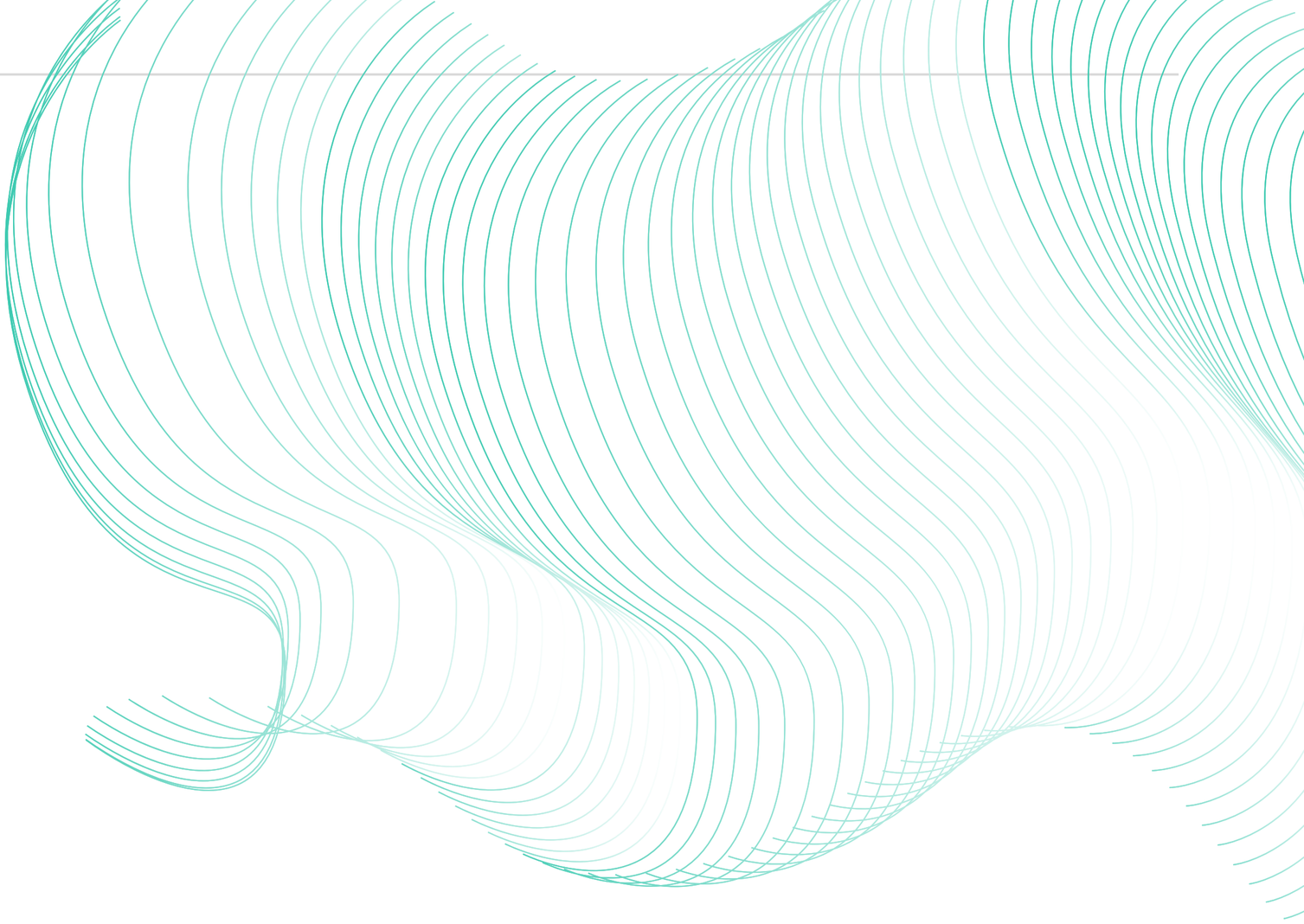
- Each year, there are approximately 6 million car accidents in the U.S.
- About 3 million people are injured every year in car accidents, with 2 million of those experiencing permanent injuries.
- Over 36,000 people died in traffic crashes in 2019.

Improved Insurance Policy Pricing

- Use of data to facilitate precision in policy pricing
- Reduction in financial risks for the insurance company
- A balanced scale, representing equilibrium in pricing and risk

Scenario

- Insurance company uses the model for new accident assessments
- Model reviews historical data to categorize accident severity
- Classification informs resource allocation for claims processing
- Enables accurate overall cost estimations
- Outcomes:
  - Efficient resource allocation
  - Enhanced customer satisfaction
  - Improved financial management



Data Understanding

US Accidents Dataset (2016-2023)

- **Source: Kaggle**
- Uploaded by: Sobhan et al.
- Origin: Paper titled "Accident Risk Prediction based on Heterogeneous Sparse Data"
- Presented: 27th ACM SIGSPATIAL International Conference, 2019.

Topic: Car accident information

Coverage: 49 U.S. states

Time Period: February 2016 - March 2023

Data Collection: Multiple APIs from sources like:

- Government departments
- Law enforcement
- Traffic cameras
- Road network sensors

Volume: 7.7 million accident records

Dataset Summary

- Car accident:
  - Each instance in the data set represents one car accident
- Total Examples:
  - 7,728,394
- Variables:
  - Total: 46
- Types:
  - Object (strings), int64, float64, boolean
- Classification Problem Insight:
  - Dominant class ratio: 0.79

Distribution of Accidents by Severity

Level 1 - Minor

- Negligible damage or injuries
- Minor cuts, bruises
- Minimal property impact

Level 2 - Moderate

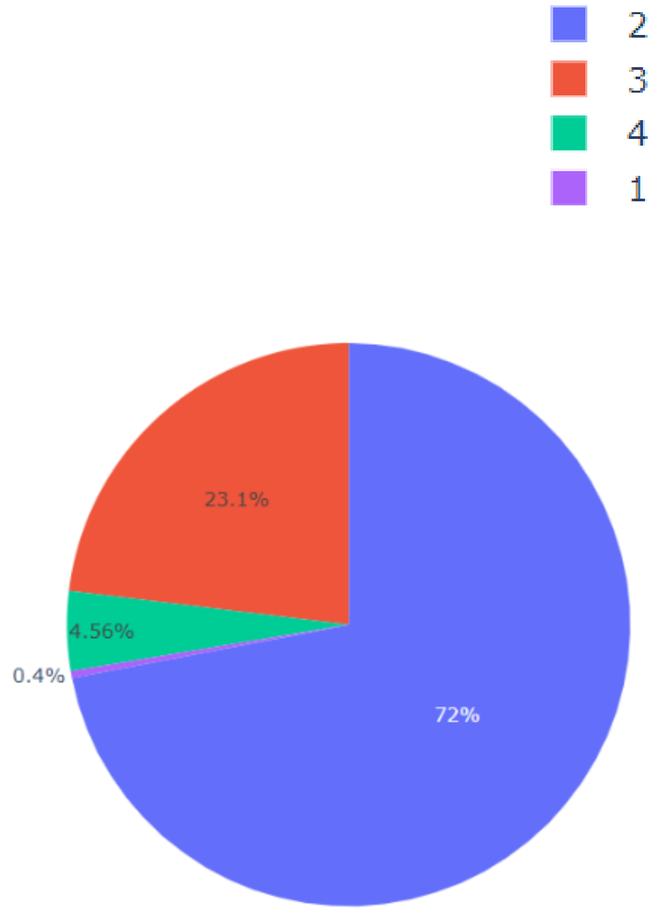
- Moderate damage and injuries
- Fractures, moderate burns
- Property repair needed

Level 3 - Serious

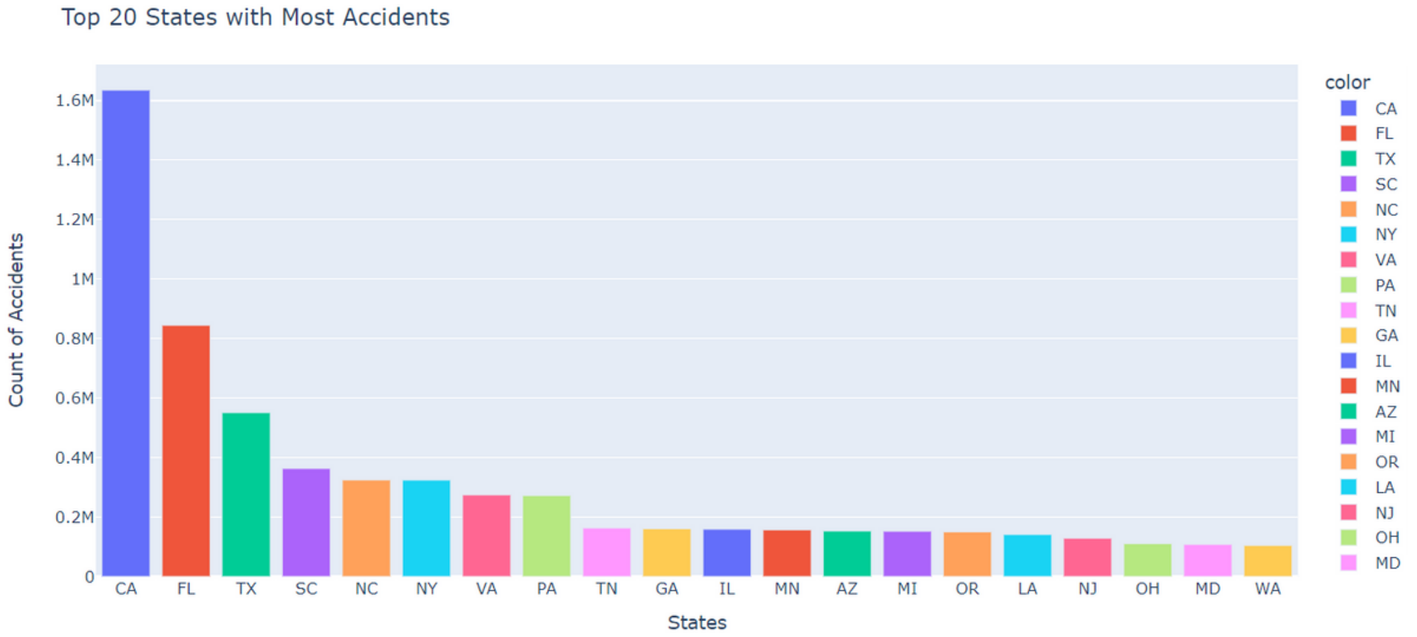
- Serious and life-threatening injuries
- Head injuries, major burns
- Significant property and financial impact

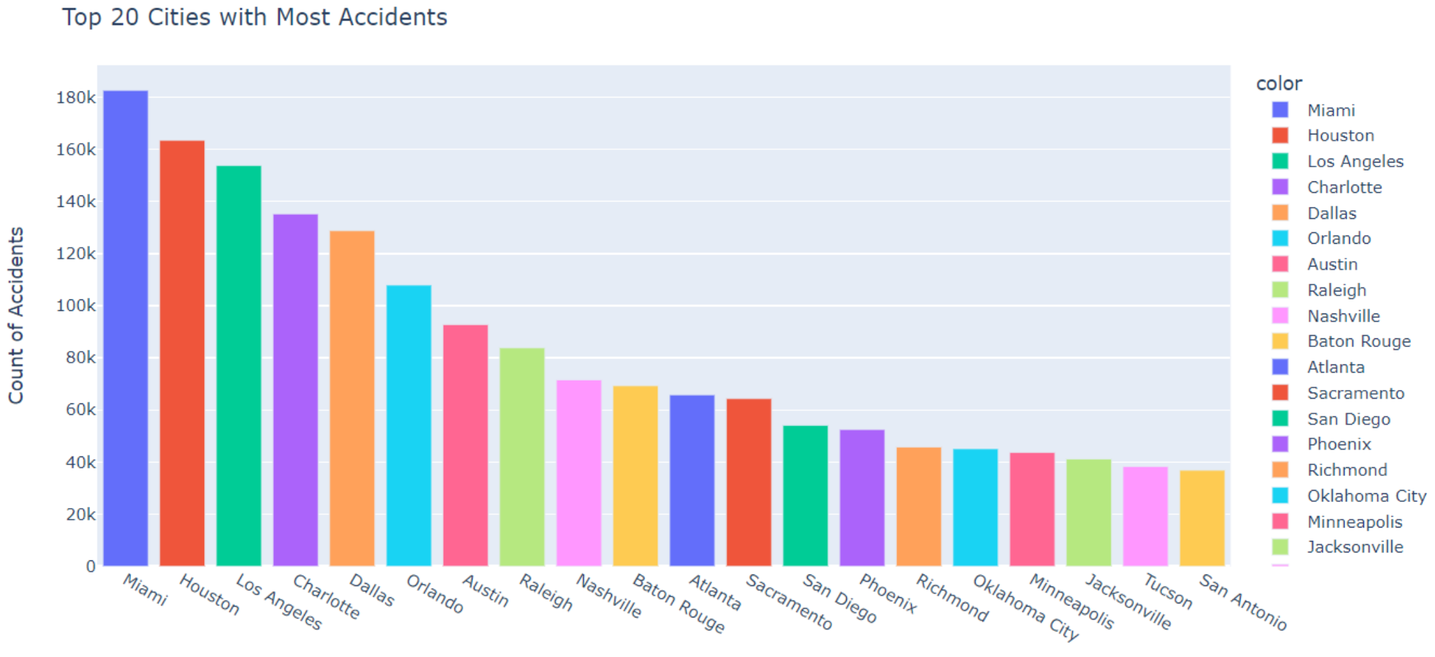
Level 4 - Catastrophic

- Catastrophic injuries or fatalities.
- Permanent disabilities.
- Widespread property damage.

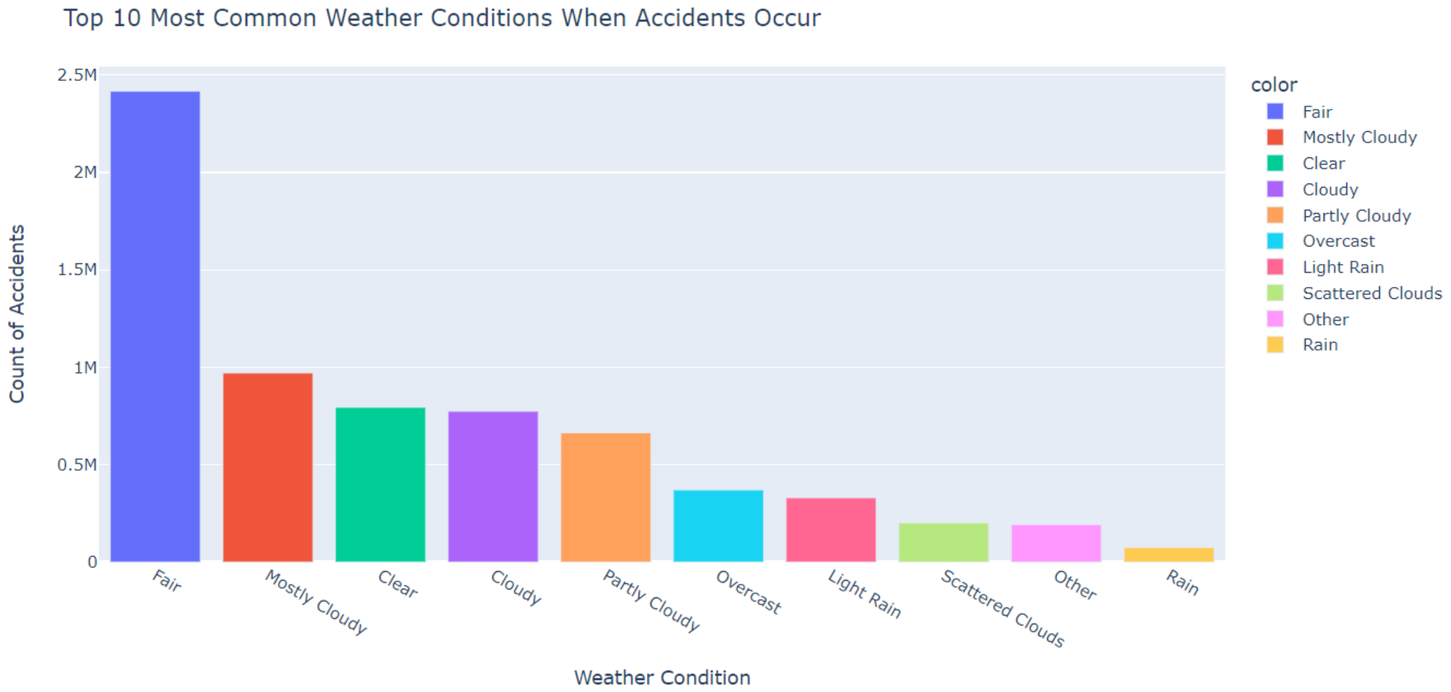


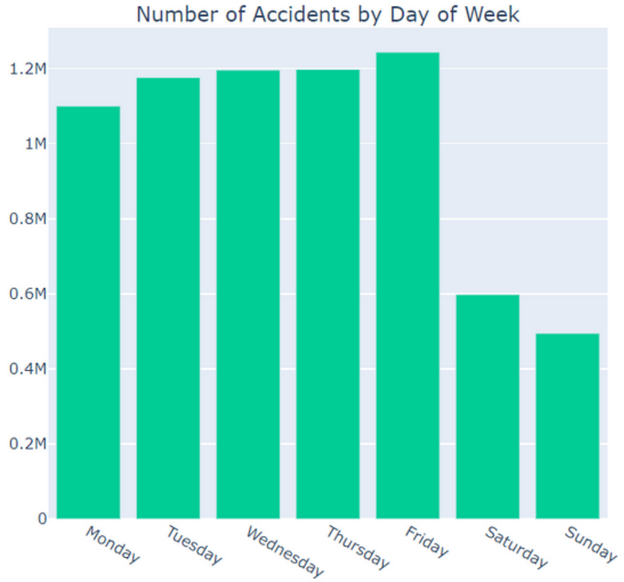
Exploratory Analysis

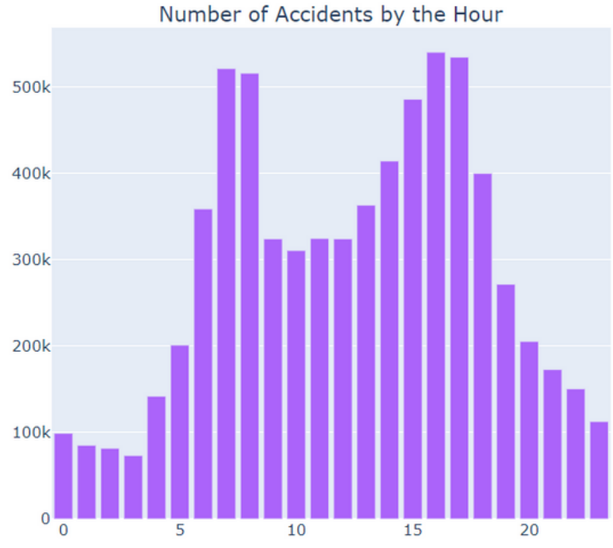




Exploratory Analysis







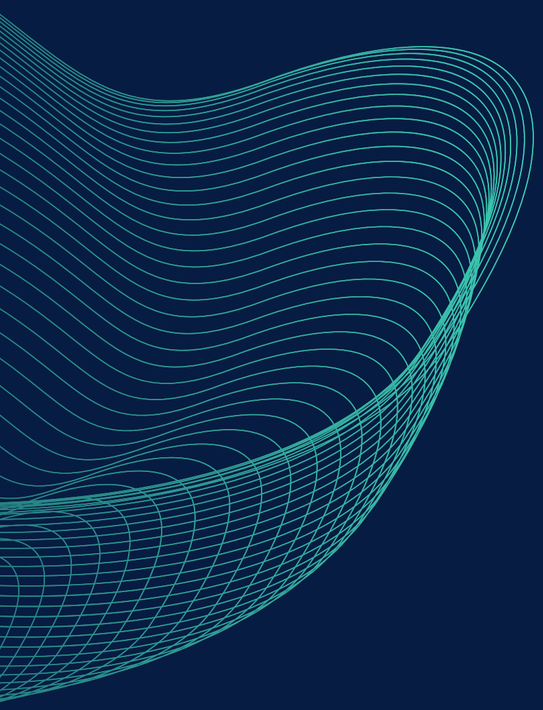
Data Preparation

1. Categorical Variables Handling:
   - Action: Converted categorical variables to dummy/indicator variables.
     - Facilitate numerical computations and model compatibility.
2. Tackling Imbalanced Dataset:
   - Action: Implemented undersampling.
     - Achieve a balanced distribution, prevent bias towards dominant class.
3. Reducing Noise and Redundancy:
   - Action: Dropped columns with single or unique values, including IDs.
     - Why: Eliminate non-informative features, streamline the dataset.

Handling Missing Values:

1. Mean Imputation:
   - When: For numerical values with approximately normal distribution.
   - Why: Captures central tendency without skewing the data.
   - Usage in Our Example: Applied where data distribution is close to normal.
2. Median Imputation:
   - When: For skewed distributions or presence of outliers.
   - Why: Median is less sensitive to outliers, preserving the data's central position.
   - Usage in Our Example: Used for variables with notable skewness or outliers

Methodology Overview



Modeling Approaches:

- Logistic Regression
- K-Nearest Neighbors (K-NN)
- Decision Tree

Validation & Hyperparameter Tuning:

- Technique: Nested Cross-Validation
  - Inner Loop: 5 folds (Hyperparameter tuning)
  - Outer Loop: 5 folds (Model performance)
- Tuning Strategy: Grid Search

Evaluation Metric:

- F1-macro (for multiclass classification)

Data Preparation:

- Standardization of numeric & ordinal features

Model Evaluation

	Algorithm	CV F1
0	Untuned KNN	0.68247
1	Untuned Decision Tree	0.74224
2	Untuned Logistic Regression	0.51798

```
Classifier                mean       +/- STD
------------------------  --------  ----------
k-NN                      0.834322  0.00537364
Decision Tree             0.866313  0.00853559
Logistic Regression       0.852351  0.00783157
```

```
 Parameter Tuning for KNN
Non-nested CV F1:  0.8362642504953742
Optimal Parameter:  {'knn__algorithm': 'auto', 'knn__n_neighbors': 13, 'knn__weights': 'uniform'}
Optimal Estimator:  Pipeline(steps=[('sc', StandardScaler()),
                ('knn', KNeighborsClassifier(n_neighbors=13))])
Nested CV F1:  0.8343217815746385  +/-  0.005373639736878231
```

```
Parameter Tuning for Decision Tree
Non-nested CV F1-score:  0.86631254342583
Optimal Parameter:  {'max_depth': 8, 'min_samples_leaf': 2}
Optimal Estimator:  DecisionTreeClassifier(max_depth=8, min_samples_leaf=2, random_state=42)
Nested CV F1-score:  0.86631254342583  +/-  0.008535593543242296
```

```
 Parameter Tuning Logistic Regression
Non-nested CV F1:  0.8523506279595289
Optimal Parameter:  {'C': 100, 'multi_class': 'auto', 'penalty': 'l2'}
Optimal Estimator:  LogisticRegression(C=100, random_state=42, solver='liblinear')
Nested CV F1: 0.8523506279595289  +/-  0.00783157208925019
```

Evaluation: Accurate risk evaluation & pricing to prevent financial setbacks

Projected Benefits:

1. Financial Savings:
    - Reduction in mispriced policies leading to potential cost savings.
    - Optimized resource allocation can reduce operational costs.
2. Enhanced Customer Satisfaction:
    - Faster claims processing and more accurate pricing.
    - Could lead to increased customer retention and referrals.
3. Accident Prevention:
    - Identifying patterns can lead to better preventive strategies.
    - Reduction in accidents can result in fewer claims and increased savings.

Challenges in Quantifying ROI:

- Intangible benefits: Brand reputation, customer trust.
- Long-term results may be gradual and not immediate.

Deployment

- Integration: Integrate model into the insurance company's claim processing system.
- Usage: Automatically assess and categorize new accident data for precise policy pricing and resource allocation.

Risks & Mitigation:

- Over-reliance: Diversify decision-making, avoid sole reliance on model results.
- Model Transparency: Ensure stakeholders understand the model's workings and limitations.
- Regular Updates: Regularly update the model with fresh data to maintain accuracy and relevance.

Potential Issues:

- Data Continuity: Ensure consistent data format for new accidents.
- Model Drift: Model's performance may degrade over time as accident patterns change.

Ethical Considerations:

- Data Privacy: Safeguard personal details associated with accident data.
- Bias and Fairness: Ensure the model does not unintentionally favor or discriminate against certain groups or types of accidents.

Recommendation

1. Adjust Premium Pricing:

Insurance companies can use the severity data to adjust premium pricing in regions or cities with a higher frequency of severe accidents (such as Los Angeles or Dallas).

2. Launch safety programs or offer discounts to policyholders in cities or states with high accident rates. Encourage safe driving behaviors through incentives and education.

3. Allocate more resources to a city/ state with a high likelihood of accidents

3. Make predictions of accident severity using our model

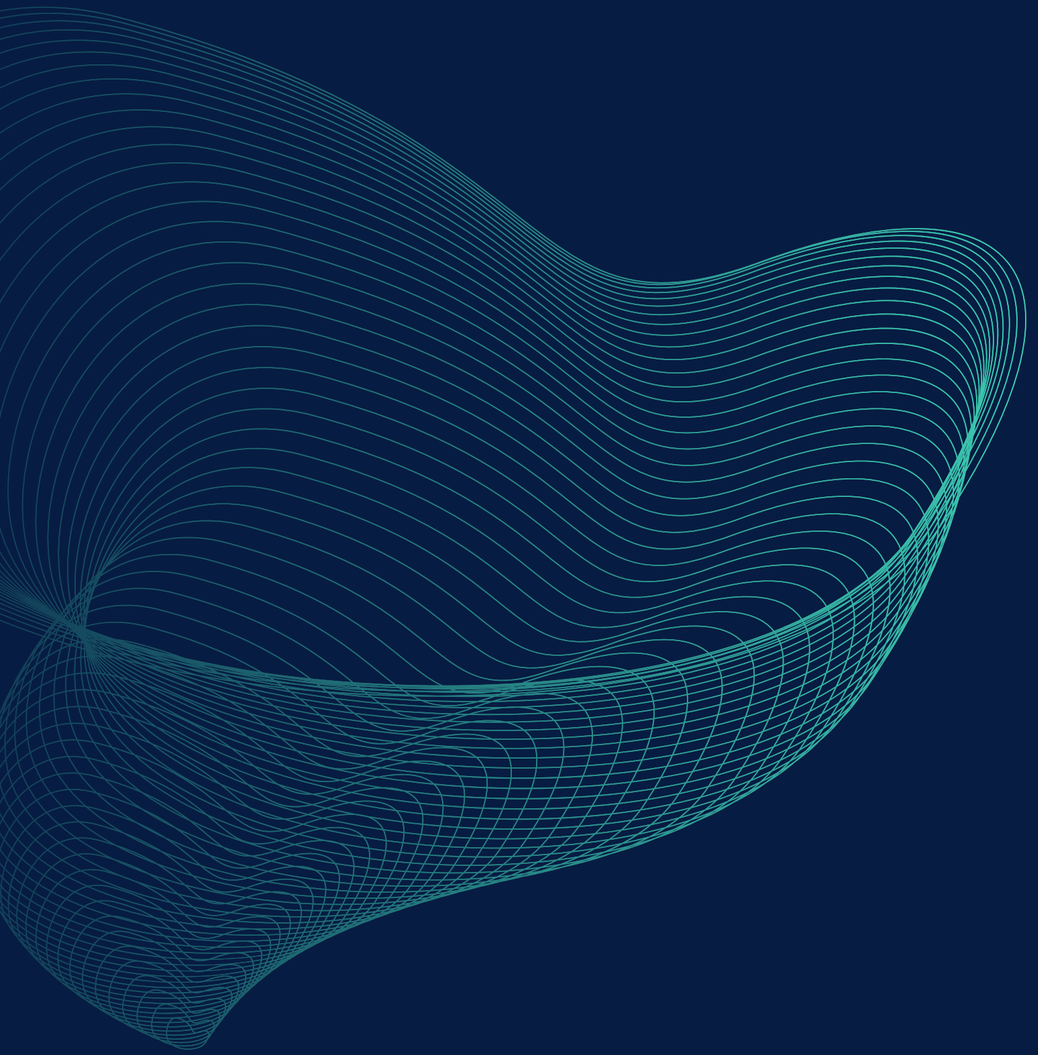
Challenges faced

1. Large Dataset - over 7 million rows

2. Limited Compute resource

3. Multiclass classification

1.



Thank you!

DataMaven