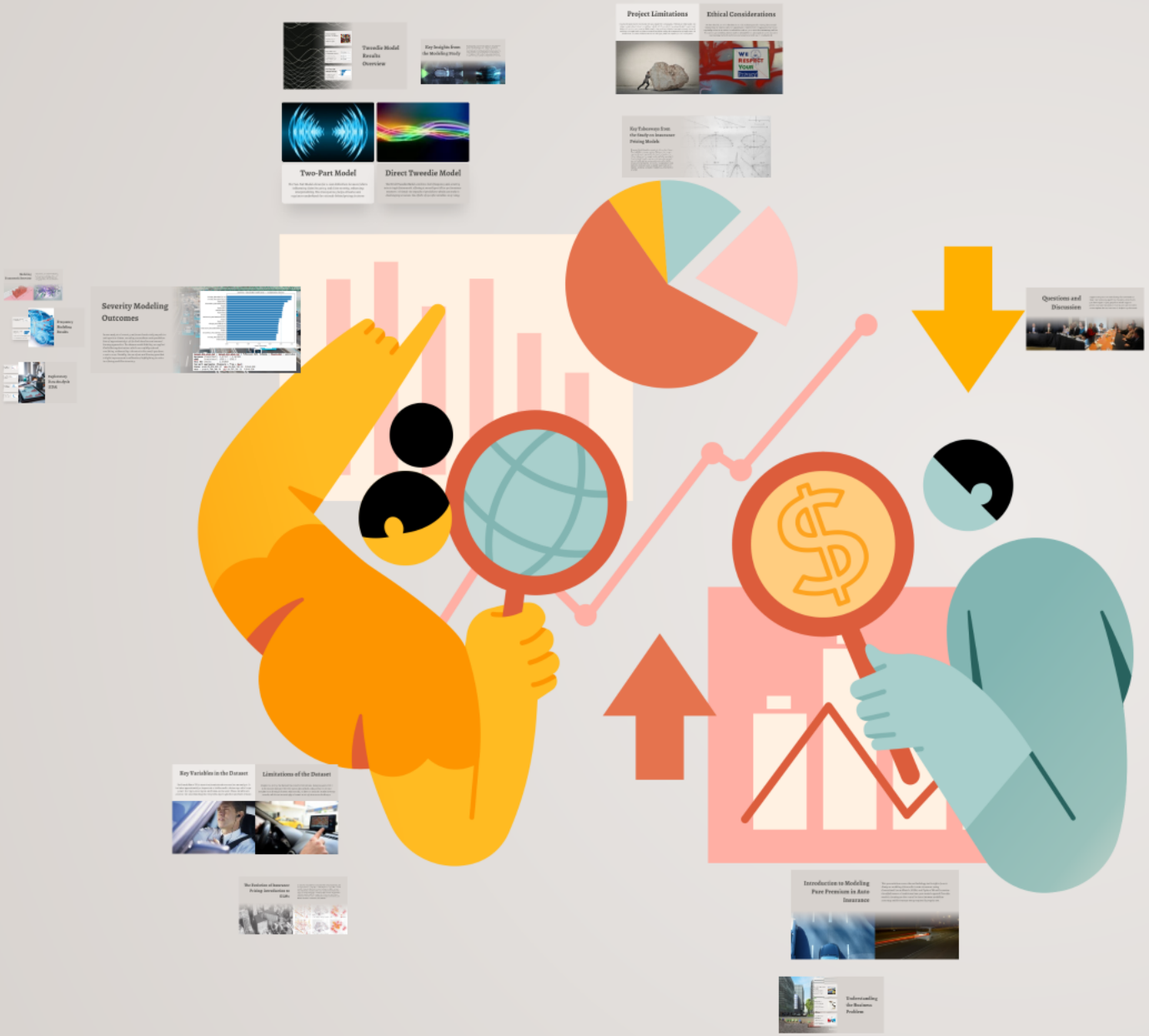


Modeling Pure Premium in Auto Insurance Using GLMs and Python

Exploring the use of Generalized Linear Models (GLMs) and Python to accurately model and predict pure premiums in auto insurance, balancing accuracy and explainability.

Introduction to Modeling Pure Premium in Auto Insurance

This presentation covers the methodology and insights from a study on modeling claim costs in auto insurance using Generalized Linear Models (GLMs) and Python. We will examine the effectiveness of traditional two-part models against Tweedie models, focusing on the crucial balance between predictive accuracy and the transparency required by regulators.

Predicting Future Costs

Insurance pricing hinges on accurately forecasting a policyholder's future claims cost. Pricing inaccuracies can lead to significant financial consequences, either through loss of customers or financial losses due to underpricing.

Regulatory Requirements

Regulatory bodies mandate that insurance models maintain transparency and fairness. Insurers must justify the reasons behind the pricing decisions for different drivers or vehicles, ensuring ethical practices in pricing.

Key Research Questions

This study investigates key questions that impact insurance pricing strategies, focusing on identifying significant characteristics, the role of segmentation, the performance comparison between modeling approaches, dataset biases, and potential improvements over basic models.

Understanding the Business Problem

Predicting Future Costs

Insurance pricing hinges on accurately forecasting a policyholder's future claims cost. Pricing inaccuracies can lead to significant financial consequences, either through loss of customers or financial losses due to underpricing.

Regulatory Requirements

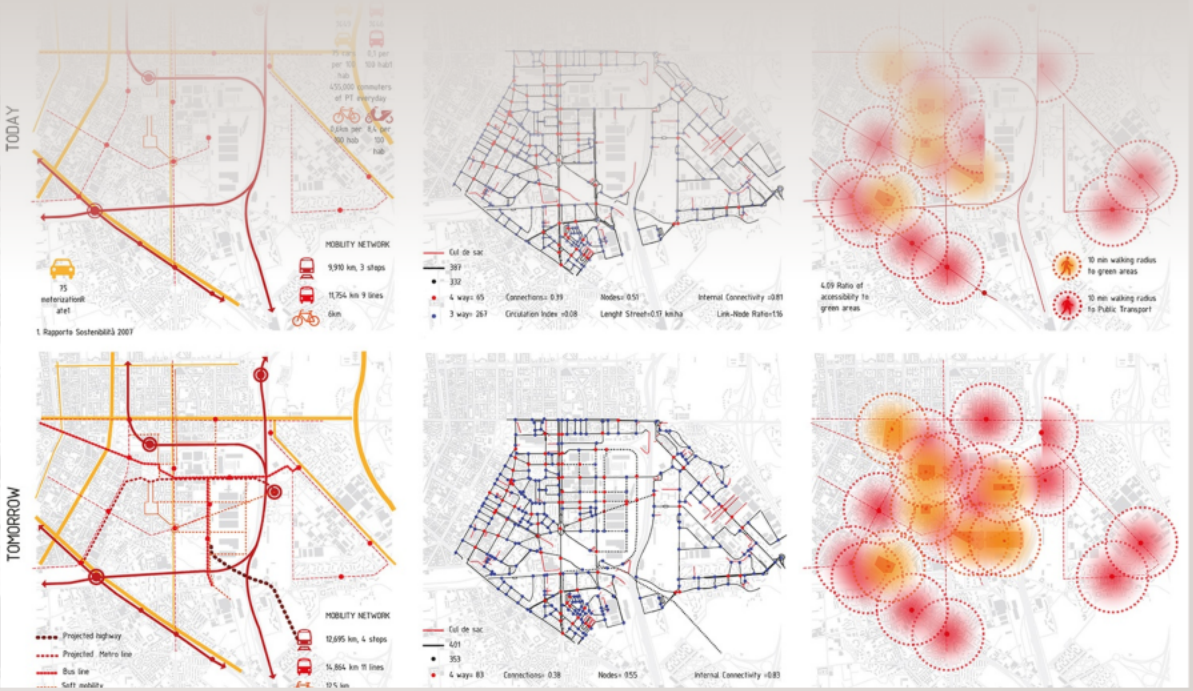
Regulatory bodies mandate that insurance models maintain transparency and fairness. Insurers must justify the reasons behind the pricing decisions for different drivers or vehicles, ensuring ethical practices in pricing.

Key Research Questions

This study investigates key questions that impact insurance pricing strategies, focusing on identifying significant characteristics, the role of segmentation, the performance comparison between modeling approaches, dataset biases, and potential improvements over basic models.

The Evolution of Insurance Pricing: Introduction to GLMs

In traditional insurance pricing, individuals were categorized broadly, such as all young drivers or urban drivers facing higher premiums. This method provided basic pricing but lacked nuance and accuracy. As data became more sophisticated, actuaries adopted Generalized Linear Models (GLMs), which offer greater flexibility in modeling. GLMs allow for the application of distributions that reflect the complexities of insurance data: Poisson distribution for claim counts, Gamma distribution for claim amounts, and Tweedie distribution for modeling pure premium.



Key Variables in the Dataset

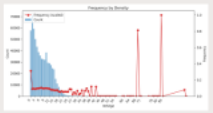
The French Motor TPL dataset is a foundational resource for our analysis. It includes approximately 10 important variables such as driver age, vehicle age, power, fuel type, area, region, and bonus–malus score. These variables are essential for understanding the risk profile of policyholders and their vehicles.

Limitations of the Dataset

Despite its utility, the dataset has notable limitations. Being 20 years old, it lacks modern features like telematics data which could provide real-time insights into driving behavior. Additionally, it does not include detailed driving records, which are increasingly relevant in today's insurance landscape.

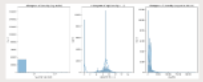
Exploratory Data Analysis (EDA)

Zero Claims Dominance



A significant portion of policies exhibit zero claims, indicating a strong zero-inflated distribution, which is a common characteristic in insurance datasets. This pattern suggests many policyholders do not file claims during the policy period.

Long-Tailed Claim Amounts

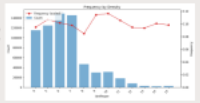


Claim amounts follow a long-tailed distribution, where the majority of claims are small; however, a few claims can be significantly large. This property of the dataset necessitates careful modeling to capture the risk accurately.

Non-Linear Relationships

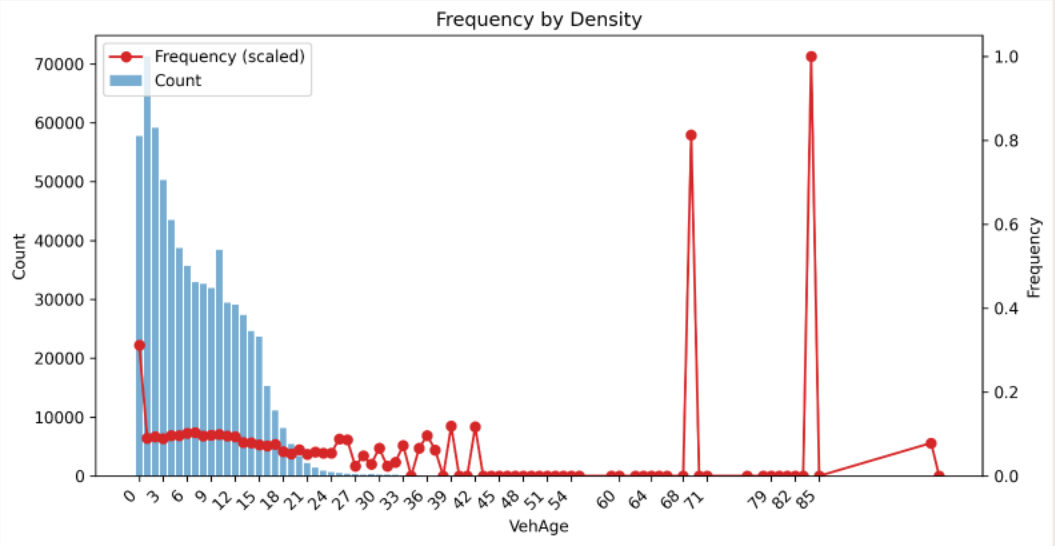
Clear non-linear relationships between age and vehicle age were identified, indicating that these factors affect pure premium in complex ways. This necessitated the creation of manual bins to enhance model interpretability and stability when analyzing results.

Outlier Management and Binning



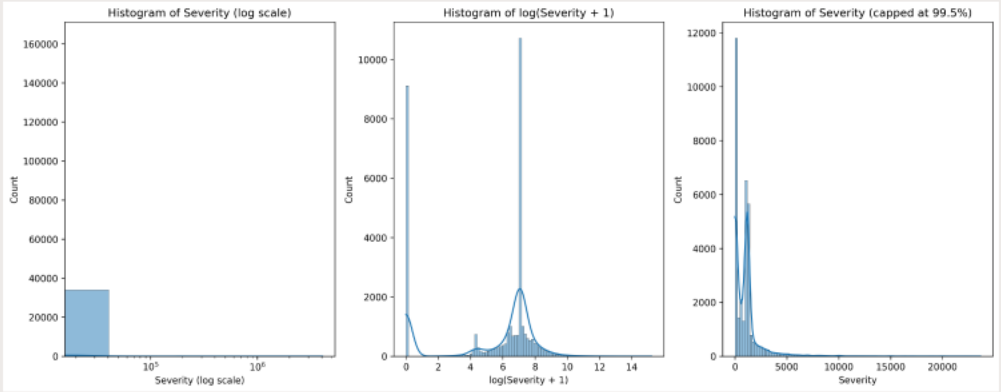
To address the challenges posed by outliers, extreme values were capped, and manual bins were created, particularly for age. This process enhances both the stability of the model and the interpretability of results, making them easier to communicate to stakeholders.

Zero Claims Dominance



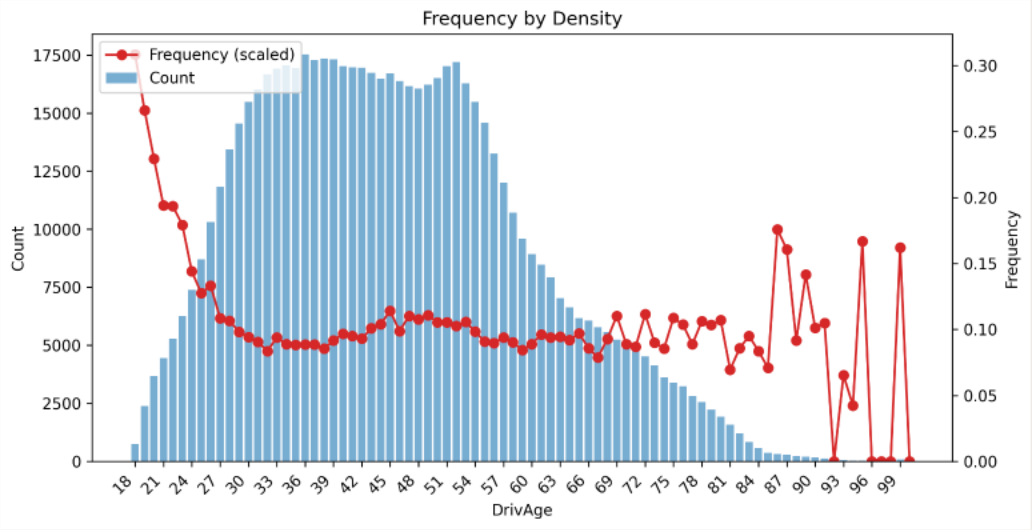
A significant portion of policies exhibit zero claims, indicating a strong zero-inflated distribution, which is a common characteristic in insurance datasets. This pattern suggests many policyholders do not file claims during the policy period.

Long-Tailed Claim Amounts



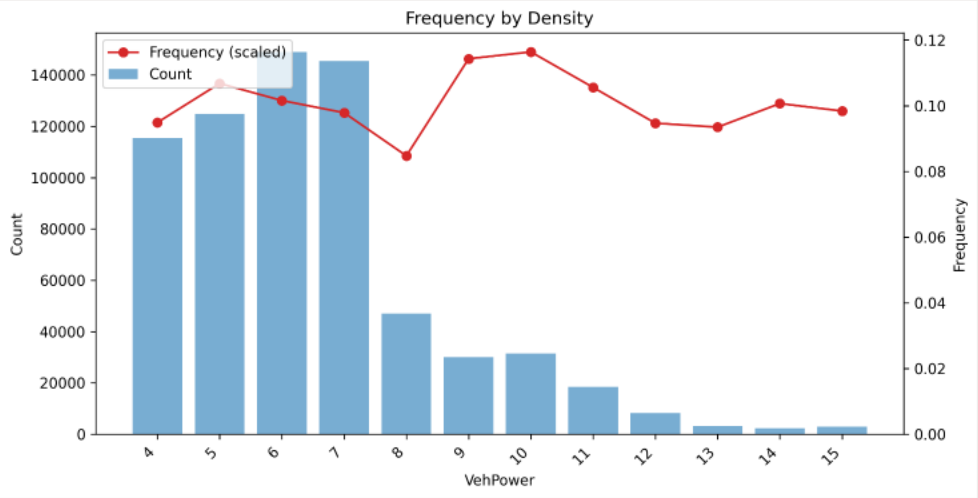
Claim amounts follow a long-tailed distribution, where the majority of claims are small; however, a few claims can be significantly large. This property of the dataset necessitates careful modeling to capture the risk accurately.

Non-Linear Relationships



Clear non-linear relationships between age and vehicle age were identified, indicating that these factors affect pure premium in complex ways. This necessitated the creation of manual bins to enhance model interpretability and stability when analyzing results.

Outlier Management and Binning

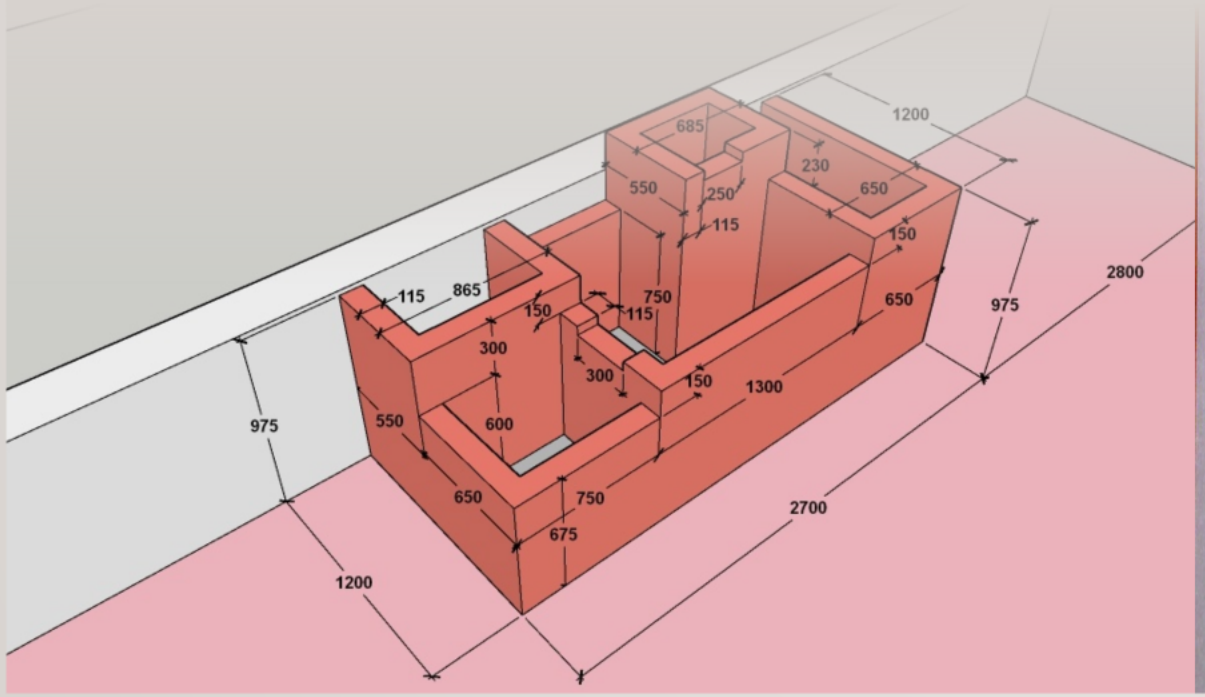


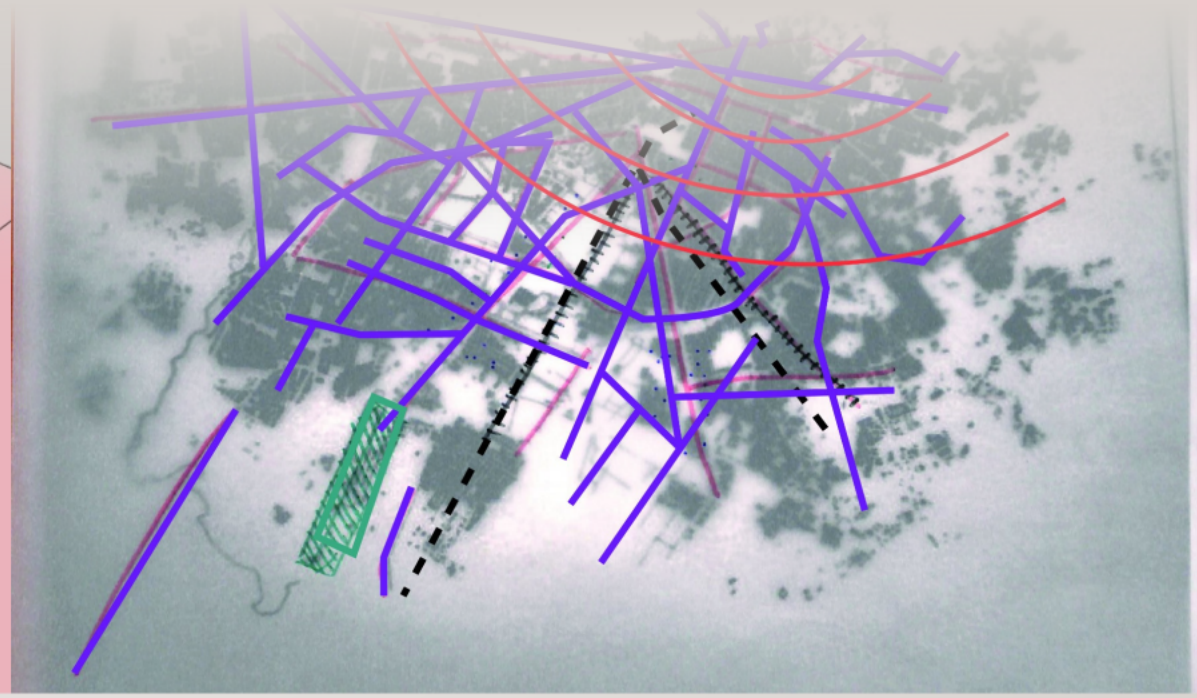
To address the challenges posed by outliers, extreme values were capped, and manual bins were created, particularly for age. This process enhances both the stability of the model and the interpretability of results, making them easier to communicate to stakeholders.

Modeling Framework Overview

Modeling Framework Overview

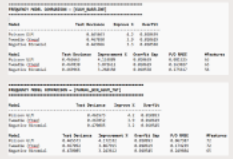
Our modeling framework comprises two distinct approaches to effectively estimate pure premium in auto insurance: the Two-Part Model and Direct Pure Premium Modeling utilizing Tweedie. The Two-Part Model employs a Poisson GLM to predict frequency and a Gamma GLM for severity. In contrast, the Tweedie model integrates both dimensions into a single framework. Each approach's performance was assessed through metrics such as deviance, mean absolute error, and enhancements compared to a null model baseline.





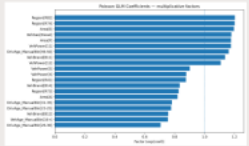
Frequency Modeling Results

Improved Model Performance with Manual Binning



In frequency modeling, we observed significant improvements with manual binning over baseline methods. Deviance reduced to 0.4656 down from 0.4657, illustrating enhanced model fit. Furthermore, the Root Mean Square Error (RMSE) improved from 0.081 to 0.067, highlighting more accurate predictions. Although these changes may seem minor, in the insurance context, even a small improvement can translate into substantial financial gains.

Segmentation Enhances Interpretability



The results underscore the importance of segmentation in frequency modeling. By utilizing manual bins, we not only achieved better stability in the model's predictions but also enhanced interpretability, making it easier for actuaries and regulators to understand the factors driving claim frequency.

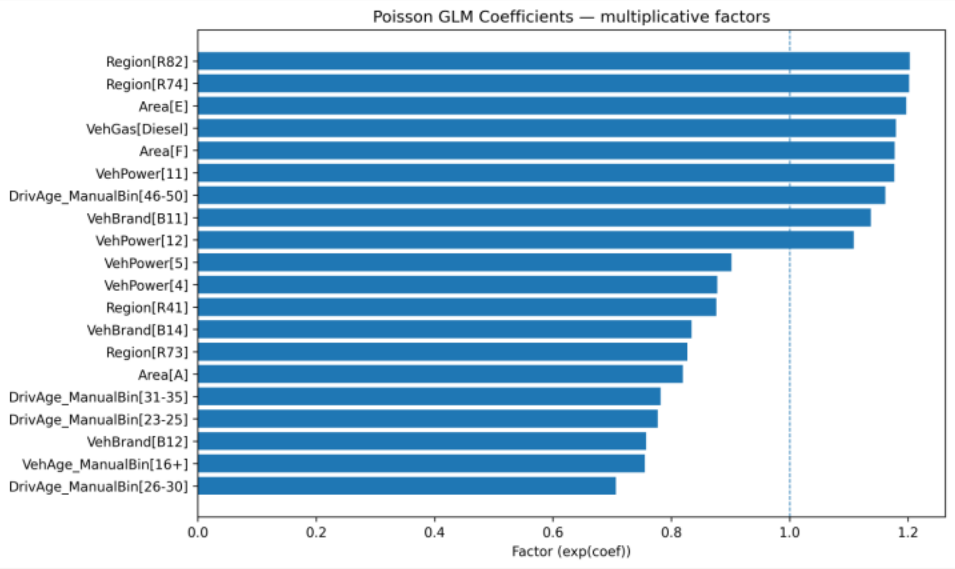
Improved Model Performance with Manual Binning

```
============================================================
FREQUENCY MODEL COMPARISON — [GLUM_BASELINE]
============================================================

Model                          Test Deviance    Improve %    Overfit
------------------------------------------------------------------------
Poisson GLM                         0.465663          4.2   0.010639
Tweedie (freq)                      0.467030          3.9   0.010629
Negative Binomial                   0.469966          3.3   0.010558

Model              Test Deviance   Improvement %   Overfit Gap   P/O RMSE   #Features
Poisson GLM             0.465663       4.153809      0.010639   0.081325          63
Tweedie (freq)          0.467030       3.872612      0.010629   0.163027          63
Negative Binomial       0.469966       3.268150      0.010558   0.275617          56
```

```
============================================================
FREQUENCY MODEL COMPARISON — [MANUAL_BIN_GLUM_TGT]
============================================================

Model                          Test Deviance    Improve %    Overfit
------------------------------------------------------------------------
Poisson GLM                         0.465573          4.2   0.010853
Tweedie (freq)                      0.467052          3.9   0.010619
Negative Binomial                   0.470085          3.2   0.010545

Model              Test Deviance   Improvement %   Overfit Gap   P/O RMSE   #Features
Poisson GLM             0.465573       4.172332      0.010853   0.067387          72
Tweedie (freq)          0.467052       3.867915      0.010619   0.174239          72
Negative Binomial       0.470085       3.243612      0.010545   0.249604          65
```

In frequency modeling, we observed significant improvements with manual binning over baseline methods. Deviance reduced to 0.4656 down from 0.4657, illustrating enhanced model fit. Furthermore, the Root Mean Square Error (RMSE) improved from 0.081 to 0.067, highlighting more accurate predictions. Although these changes may seem minor, in the insurance context, even a small improvement can translate into substantial financial gains.

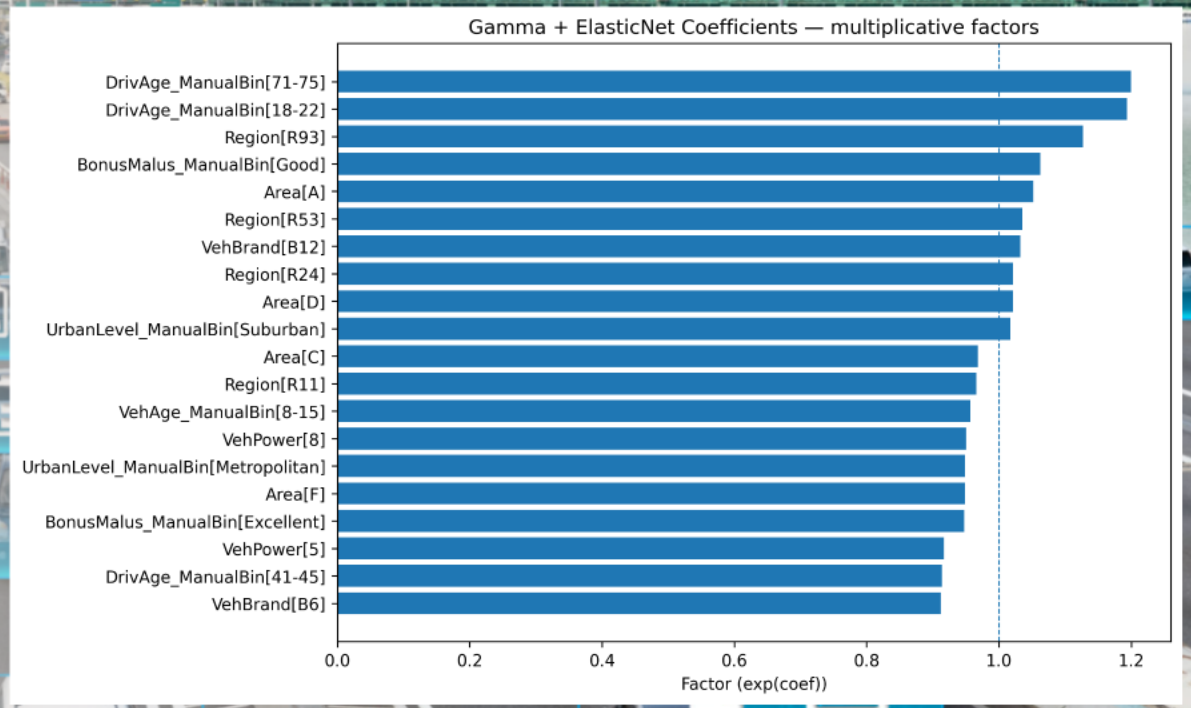
Segmentation Enhances Interpretability



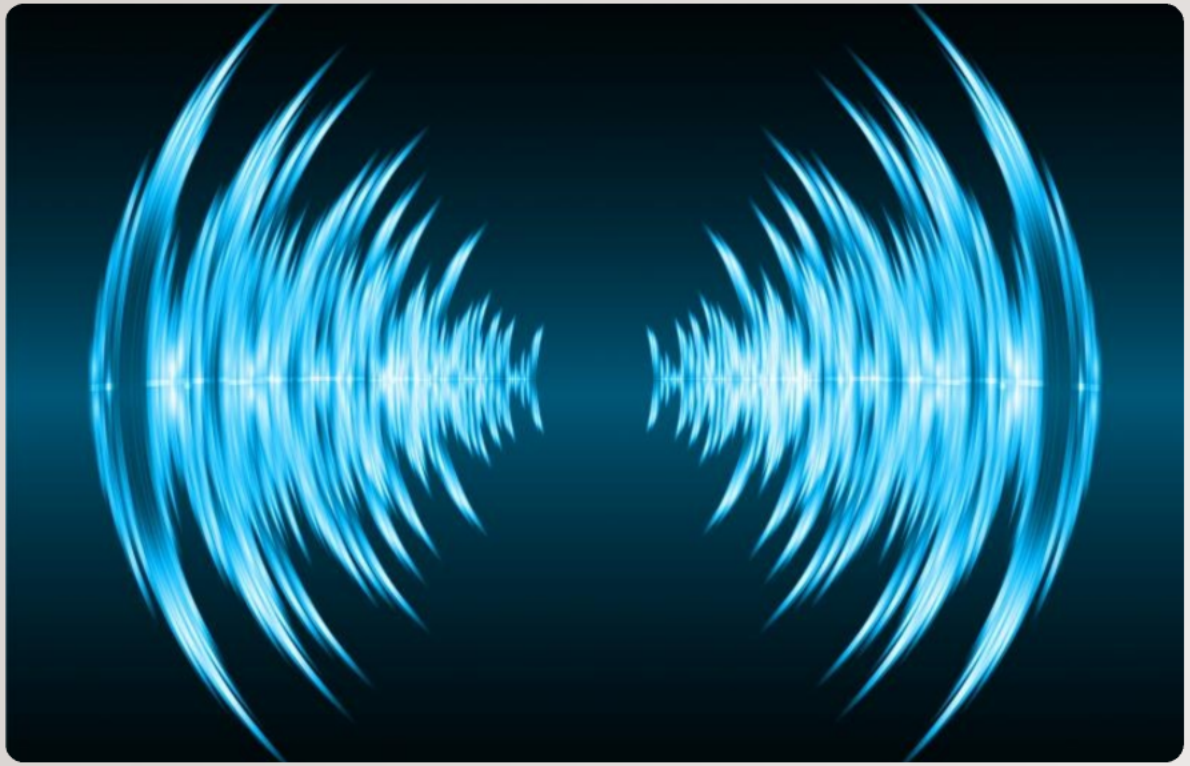
The results underscore the importance of segmentation in frequency modeling. By utilizing manual bins, we not only achieved better stability in the model's predictions but also enhanced interpretability, making it easier for actuaries and regulators to understand the factors driving claim frequency.

Severity Modeling Outcomes

In our analysis of severity, we focused exclusively on policies with positive claims, revealing a consistent underprediction bias of approximately 6.5% for both baseline and manual binning approaches. To enhance model stability, we applied ElasticNet regularization, which successfully reduced overfitting, evidenced by a decrease in the overfit gap from 0.096 to 0.091. Notably, the use of manual binning provided a slight improvement in calibration, highlighting its value in refining predictive accuracy.




```
manual_bin_glum_tgt × manual_bin_glum_tgt | F=Poisson GLM, S=Gamma + ElasticNet — pointwise
deviance (train/test): 1.277417 / 1.367991
wMAE     (train/test): 1529.1 / 1592.5
null dev (test):       1.371625
two-part aggregates (Exposure × Freq × Sev):
train: pred=38,846,092.57  obs=38,869,709.33  P/O=0.999
test : pred=9,704,108.50  obs=10,387,183.21  P/O=0.934
```



Two-Part Model

The Two-Part Model allows for a clear distinction between factors influencing claim frequency and claim severity, enhancing interpretability. This transparency helps actuaries and regulators understand the rationale behind pricing decisions.

Direct Tweedie Model

The Direct Tweedie Model combines both frequency and severity into a single framework, offering a smooth predictive performance. However, it blends the impacts of predictors, which can make it challenging to isolate the effects of specific variables on pricing.

Unified Modeling of Frequency and Severity

Tweedie models provide a unified approach to modeling both frequency and severity, allowing insurers to predict pure premiums more effectively. The power parameter is crucial as it adjusts the balance between treating the response variable as a count (frequency) or a continuous variable (severity).

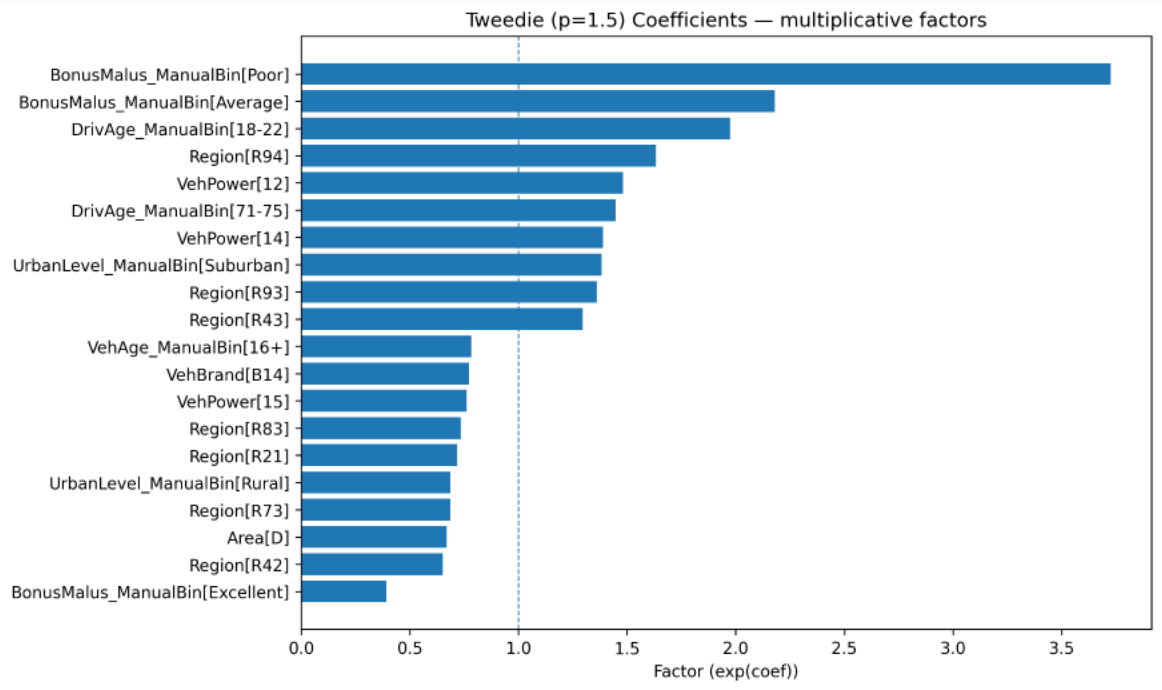
Optimal Tweedie Performance Metrics

At a power parameter of p = 1.5, the Tweedie model achieved a deviance of 76.63, indicating it performed better than other model configurations. Furthermore, it demonstrated a lower underprediction rate of 5.1%, enhancing its reliability in estimating pure premiums.

```
============================================================
PURE PREMIUM — glum_baseline_pp
============================================================

Tweedie (p=1.9) (p=1.899999976158142)
training loss (deviance): 33.546899
testing  loss (deviance): 33.885491
train totals  observed = 38,869,709.33, predicted = 40,735,162.56
test  totals  observed = 10,387,183.21, predicted = 10,173,501.52

Tweedie (p=1.7) (p=1.7000000476837158)
training loss (deviance): 35.900331
testing  loss (deviance): 37.002309
train totals  observed = 38,869,709.33, predicted = 39,835,183.03
test  totals  observed = 10,387,183.21, predicted = 9,954,784.11

Tweedie (p=1.5) (p=1.5)
training loss (deviance): 73.021887
testing  loss (deviance): 76.633913
train totals  observed = 38,869,709.33, predicted = 39,295,679.83
test  totals  observed = 10,387,183.21, predicted = 9,822,788.68

Tweedie (p=1.3) (p=1.2999999523162842)
training loss (deviance): 189.789315
testing  loss (deviance): 203.048486
train totals  observed = 38,869,709.33, predicted = 38,998,966.52
test  totals  observed = 10,387,183.21, predicted = 9,749,653.84
```

Challenges in Interpretability

Despite its advantages, the Tweedie model's interpretability is a significant concern. Its blending of frequency and severity makes it challenging for regulators and stakeholders to understand the specific contributions of each factor to the risk assessment.



Key Insights from the Modeling Study

The analysis revealed several critical insights regarding the predictors and modeling techniques used in auto insurance pricing. Notably, driver age and vehicle age emerged as the most significant predictors of pure premium. The use of manual binning consistently enhanced the interpretability of the models, making it easier to communicate findings to stakeholders. Additionally, the application of regularization techniques effectively reduced overfitting in severity modeling. While the Tweedie model exhibited superior predictive performance, the two-part model maintained greater explainability. Overall, improvements in deviance ranged between 3% to 5% over the null model, highlighting the efficacy of the approaches utilized in this study.

Project Limitations

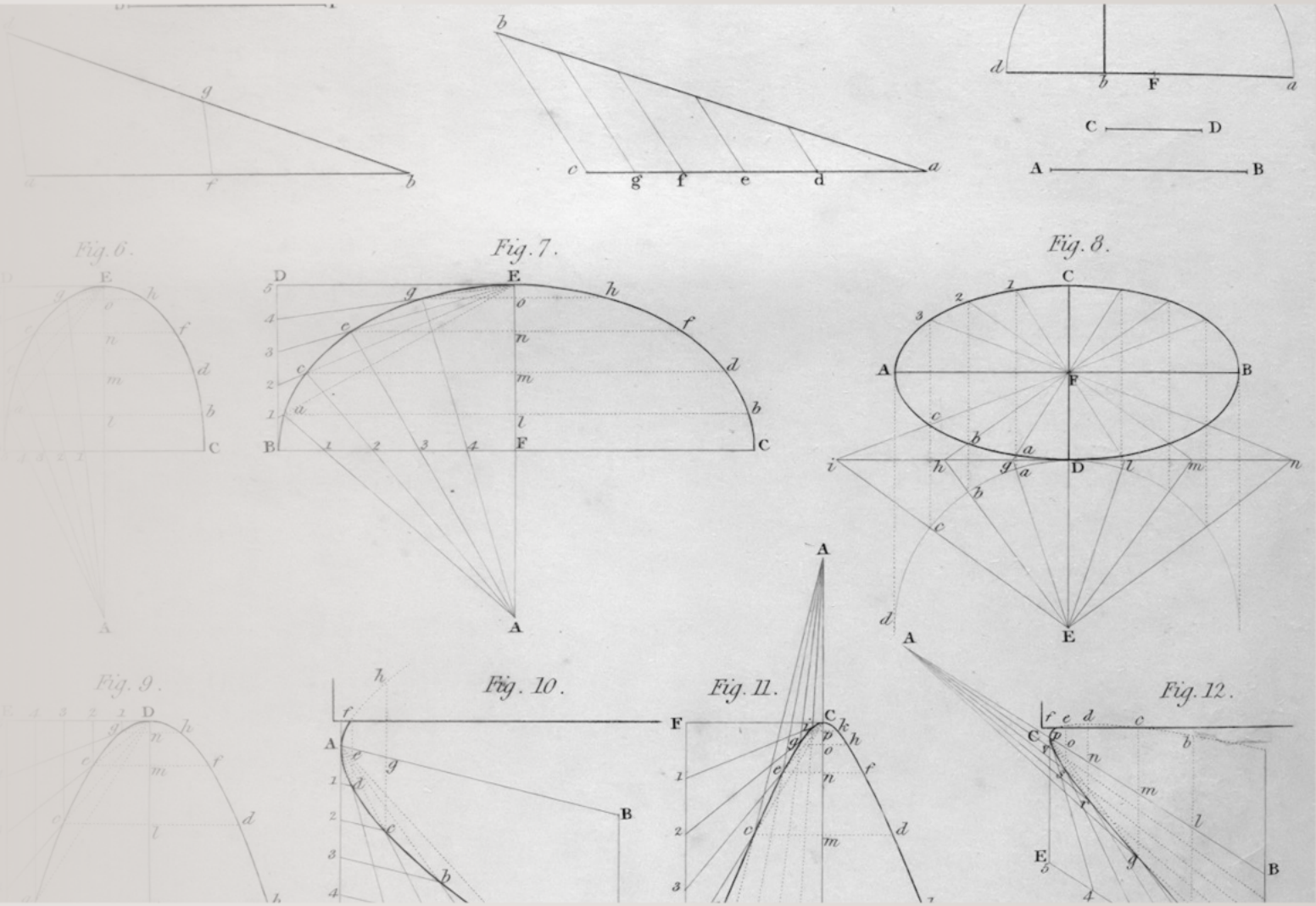
This project faced several limitations that may impact the accuracy and effectiveness of the models. The dataset contained only about 10 predictors, which restricts our ability to capture the full complexity of factors influencing pure premium. Additionally, time constraints limited our exploration of advanced modeling techniques such as splines or interaction effects. Lastly, the interpretation of coefficients and residuals in the context of GLMs remains a challenging task that requires further investigation.

Ethical Considerations

On the ethical front, while the dataset was old and anonymized, ensuring fairness and transparency in insurance pricing is paramount. Insurers must navigate potential biases, especially related to sensitive characteristics such as age, to avoid discriminatory practices. Transparency in modeling choices and the rationale behind pricing decisions is crucial for maintaining trust with customers and compliance with regulatory standards.

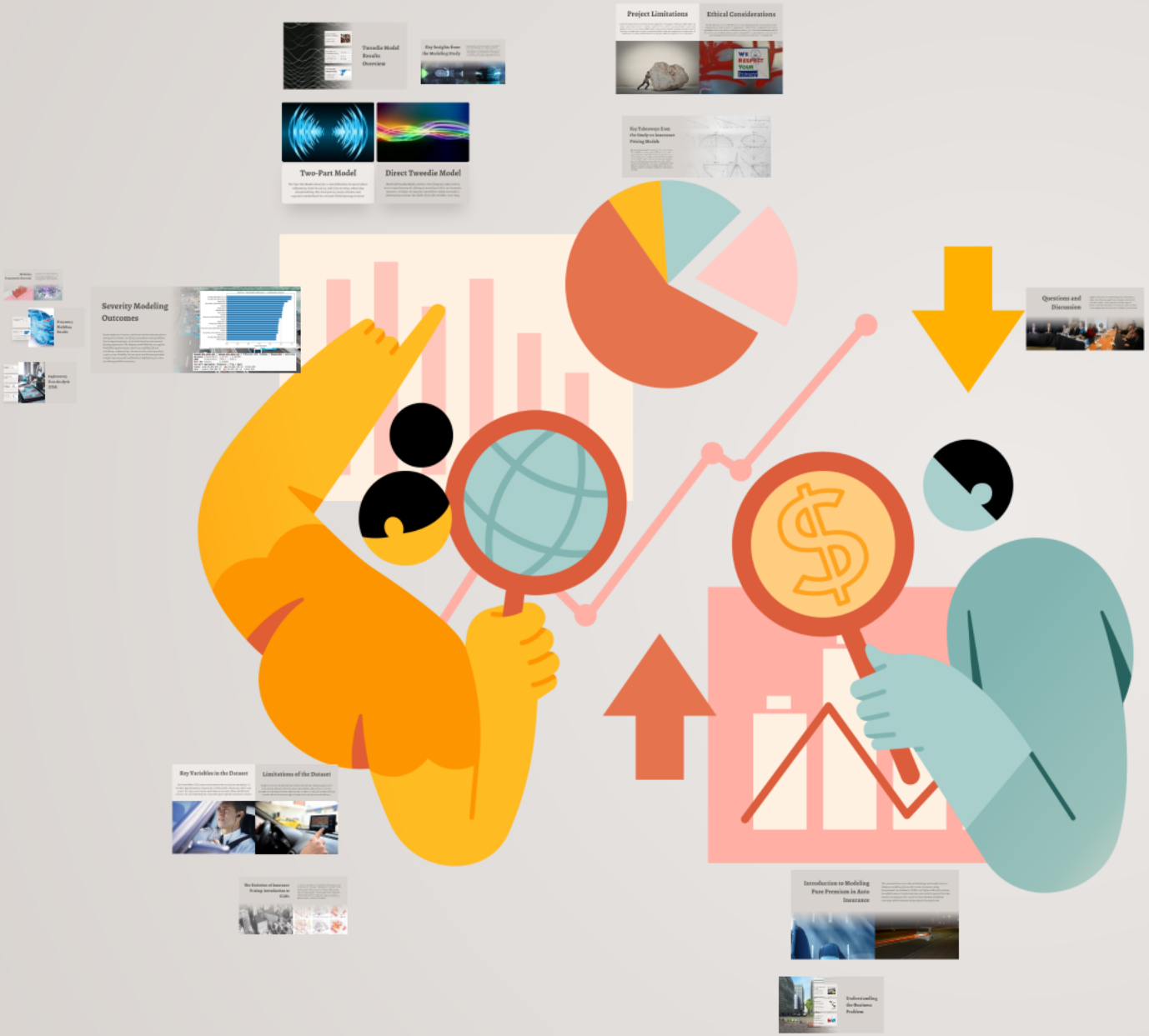
Key Takeaways from the Study on Insurance Pricing Models

This project highlighted the pivotal role of Generalized Linear Models (GLMs) in insurance pricing. We found that two-part models enhance interpretability by clearly delineating the factors influencing frequency from those affecting severity. On the other hand, Tweedie models provide a more streamlined approach with slightly superior performance, albeit at the expense of interpretability. Ultimately, the appropriate model selection is contingent upon the specific requirements of the business, balancing the need for transparency with predictive accuracy.



Questions and Discussion

I appreciate your attention during this presentation. Now, I welcome any questions, thoughts, or feedback you have regarding my project on modeling pure premium in auto insurance. Your insights are valuable as we explore the implications of the findings discussed.



Modeling Pure Premium in Auto Insurance Using GLMs and Python

Exploring the use of Generalized Linear Models (GLMs) and Python to accurately model and predict pure premiums in auto insurance, balancing accuracy and explainability.