Dynamic Pricing for Auto Rental Insurance

by

Pallavi Kandanur

A Creative Component submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Artificial Intelligence

Program of Study Committee:
Simanta Mitra, Major Professor
Gurpur M Prabhu

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation/thesis. The Graduate College will ensure this report is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2024

Copyright © Pallavi Kandanur, 2024. All rights reserved.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

ACKNOWLEDGMENTS

I would like to take this opportunity to express my thanks to Professor. Simanta Mitra and Professor Gurpur M. Prabhu for their guidance, patience, and support throughout this research and the writing of this report.

I would also like to thank Brandon Rockow, CEO of Bonzah Insurance Company for giving us the opportunity to work on the project by sharing the necessary data and industry insights.

Also , my project partner Minghao who worked with me to successfully complete the entire project.

ABSTRACT

This project focuses on developing a machine learning framework for dynamic insurance pricing in short-term rental insurance. The framework predicts accident severity and calculates personalized insurance costs by leveraging driver, location, and weather information. Using publicly available datasets and private sales and claims records, the model applies a probabilistic approach to assess accident likelihood and severity,integrating diverse feature sets through a multi-task neural network architecture and using XGBoost model. The methodology involves embedding-based feature transformations, data augmentation strategies, and advanced machine learning techniques to address challenges like data sparsity and class imbalance. Offline evaluations compare the proposed model's dynamic pricing capabilities against a static baseline, revealing significant improvements in net income, pricing accuracy, and profitability metrics. Despite data limitations, the results highlight the effectiveness of the approach in aligning insurance premiums with actual risks.

CHAPTER 1. INTRODUCTION

Historically, car insurance pricing has been determined by a limited set of fixed factors—age, driving history, and location. While these parameters help insurance companies segment drivers into risk categories, they do not take into account the ever-changing, real-time factors that influence an individual's true risk. Companies like Progressive and Allstate have used this method, where rates remain unchanged over time, but this can result in premiums that do not reflect the actual risk presented by each driver.

With the rise of new technologies, insurance pricing has become more dynamic and personalized. One of the most significant innovations is the use of telematics, which involves collecting data through GPS, accelerometers, and sensors embedded in vehicles. This technology captures detailed driving behavior, such as speed, braking intensity, and acceleration, allowing insurers to assess risk with greater precision and offer more tailored premiums based on individual driving habits . For example, Allstate's Drivewise program Saf (2024) leverages this data to adjust premiums, offering discounts to drivers who demonstrate safe habits behind the wheel, thereby promoting safer driving practices and rewarding responsible customers with lower rates. In addition, research efforts such as 'Using GPS data to analyse the distance travelled to the first accident at fault in pay-as-you-drive insurance' Ayuso et al. (2016) demonstrate the use of GPS data and driving patterns to assess accident risk, providing valuable insights for pay-as-you-drive models. These studies demonstrate the potential of incorporating real-time driving data into insurance pricing, offering more personalized and accurate rates. Despite this, current pricing models have yet to fully integrate these advancements, often relying on static, generalized factors rather than adapting rates based on the unique driving behaviors and conditions of each individual. This gap highlights the need for a more dynamic approach to pricing that can respond to real-time data and provide fairer premiums.

With this project, we aim to develop a dynamic pricing model that leverages public data including location and weather information, alongside company data, such as driver violation history, sales data, and claims records, to assign a fair, individualized premium for each driver. This model moves beyond traditional, static risk classifications, using machine learning to tailor insurance premiums based on real-time and historical data. For example, drivers in safer weather conditions or low-risk areas will receive lower premiums, while those in high-risk zones or adverse weather will see higher rates. By incorporating multiple datasets, the model evaluates and adjusts to the individual risk factors each driver presents at the time of rental.

CHAPTER 2. Background and Literature Review

As consumer expectations shift towards personalized services, the insurance industry faces a growing demand for pricing models that reflect individual behaviors rather than general categories. Research on telematics and machine learning has offered promising solutions by leveraging real-time data to create tailored premiums. This review will examine foundational and recent studies that highlight the move towards individualized car insurance pricing. One influential study by Verbelen, Antonio, and Claeskens explores how telematics data can transform insurance pricing by focusing on actual driving behaviors rather than traditional demographic factors Verbelen et al. (2018). Using data from a Belgian telematics program for young drivers, the authors demonstrate how premiums can be designed based on driving behavior collected via black-box devices installed in vehicles. Unlike traditional models, which often depend on demographic variables like age and postal code, this study applies generalized additive models (GAMs) with compositional predictors to assess the effect of telematics data—such as road types, driving times, and distance driven—on expected claim frequency. Their findings reveal that telematics-based variables significantly enhance predictive accuracy, making conventional rating factors like gender less relevant. By providing an evidence-based approach to individualized pricing, this study talks about the potential of telematics to make the shift from static factors to dynamic factors.

Pesantez-Narvaez, Guillen, and Alcañiz (2019) further explore the predictive power of telematics data by comparing XGBoost, a modern machine learning method, with traditional logistic regression for predicting motor insurance claims Pesantez-Narvaez et al. (2019). Using GPS-based data on driving behaviors such as speed, distance, and driving schedules, their study highlights a key trade-off between accuracy and interpretability in insurance pricing models. While XGBoost demonstrates high predictive accuracy, it requires extensive tuning to manage

overfitting, particularly with imbalanced datasets. Logistic regression, though less complex, provides clear interpretability, making it a robust choice for regulatory contexts that demand transparency in premium calculations. This research emphasizes how telematics, combined with carefully chosen analytical methods, can support the shift towards individualized pricing models that align premiums more closely with each driver's unique behavior, balancing precision with practical considerations in insurance regulation.

In addition to telematics based approach, machine learning has also been employed to traditional insurance datasets using demographic and claims history data to improve risk prediction and refine premium calculations.

For example, Power, Côté, and Duchesne (2024)Power et al. (2024) propose a hierarchical model that integrates XGBoost with copulas to handle dependencies within claims data. This two-part model first uses XGBoost to predict the likelihood of a claim, leveraging the algorithm's strength in handling large, complex datasets. Then, it uses copulas to model the claim amount, capturing dependencies between different factors, such as accident type and location, to improve prediction accuracy. Results show that this approach yields better accuracy and captures global data patterns more effectively than conventional generalized linear models (GLMs), making it a promising method for individualized premium calculation.

Machine Learning applications in insurance were further extended by Alamir et al. (2021) Alamir et al. (2021).Using data preprocessing techniques and K-Fold cross-validation, the authors apply Random Forest (RF) and Multi-Class SVM classifiers to categorize claims into statuses like close, pending, and settled. Random Forest achieved slightly better accuracy (98.36 percent) than SVM (98.17percent), showing the effectiveness of traditional machine learning methods in streamlining claims processing. The successful development of a machine learning model that accurately predicts the status of motor insurance claims using traditional data source will assist insurance companies in handling high claim volumes more efficiently.

Wang et al. (2020) Wang (2020) leverage machine learning to enhance auto insurance renewal predictions by identifying critical features such as insurance channel, NCD, vehicle age, and

purchase price Wang (2020). Comparing Random Forest, GBDT, and LightGBM, they conclude that LightGBM offers superior accuracy and effectively handles large insurance datasets. This research underscores the impact of feature selection and model choice in predictive insurance modeling, supporting our project's aim to use machine learning for tailored, data-driven pricing based on individual risk profiles.

A study by Xie Xie (2021) explores the potential of artificial neural networks (ANN) for predicting claim counts, claim amounts, and average loss per claim in auto insurance, aiming to refine rate-making processes. By analyzing key risk factors like accident year and coverage type, the research demonstrates how ANN can provide more accurate and interpretable results than traditional models such as GLM.The study highlights ANN's ability to capture complex relationships without strict assumptions, making it a powerful tool for precise risk assessment. This approach is closely related to our project's focus on individualized insurance pricing using machine learning, as it underscores the importance of both accuracy and transparency in setting fair and tailored premiums.

Another area of research for our project was handling imbalanced datasets. This is because at an insurance company, the drivers how put in a claim are usually a small portion of the overall drivers that opted for the rental insurance company. This was studied by Hanafy and Ming (2021)Hanafy and Ming (2021). Their research tackles the challenge of predicting auto insurance claims in highly imbalanced datasets, where claims are much fewer than non-claims. They tested various data-level techniques—under-sampling, over-sampling, hybrid methods, and SMOTE—across 32 machine learning models, including decision trees, boosting, and bagging. Their results showed that AdaBoost with oversampling or a hybrid approach provided the highest accuracy, with a sensitivity of up to 92.94 percent and specificity of 99.82 percent. This study demonstrates that adjusting for data imbalance can significantly improve claim prediction accuracy in auto insurance.

In summary, the reviewed studies collectively highlight the transformative potential machine learning in car insurance pricing. By shifting from static demographic factors to dynamic,

real-time data such as driving behavior and environmental conditions, these approaches significantly enhance predictive accuracy and personalization. Advanced models like XGBoost, LightGBM, and neural networks demonstrate superior performance in risk prediction, while methods addressing imbalanced datasets ensure robustness in claim forecasting. The integration of behavior-driven insights, efficient feature selection, and innovative data handling techniques provides a foundation for developing individualized and adaptive pricing models. These findings underscore the opportunity to create fairer, more accurate premiums, aligning closely with our project's objective of leveraging diverse data streams to implement a dynamic insurance pricing system.

CHAPTER 3. Project Framework and Implementation

This project introduces a novel approach to dynamic insurance pricing by leveraging advanced machine learning techniques and integrating multiple data streams to assess and price individual risk accurately. The core of the design lies in a step-by-step evaluation process that predicts accident likelihood, estimates severity, and calculates personalized premiums tailored to each driver's profile. (See 3.1) .The modular framework ensures adaptability, precision, and scalability in handling diverse scenarios.

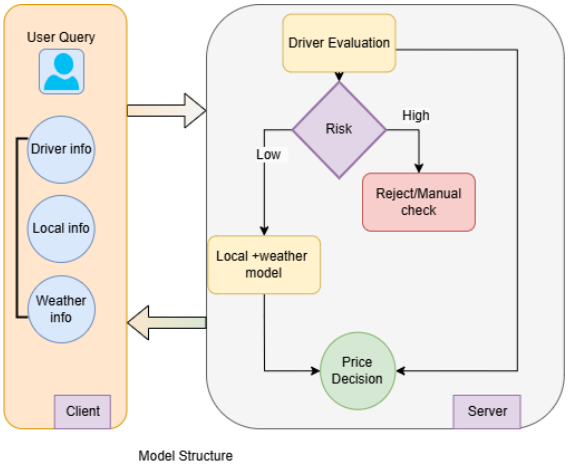

Figure 3.1 Project Design

The system is a two-stage risk evaluation process. The first stage focuses on driver-specific data, analyzing features such as violation history, age, and driving behavior to classify drivers into distinct risk categories. High-risk drivers are flagged for manual review or rejection, while low-risk drivers proceed to further evaluation. This approach ensures that extreme-risk cases are handled with caution, maintaining the reliability and accuracy of the overall system. By isolating driver risk as an initial step, the system creates a robust foundation for subsequent analysis.

The second stage uses the Local + Weather Model to refine the risk assessment by incorporating contextual and environmental factors.This is done for the low-risk drivers. This

model analyzes features such as state, time of day, temperature, humidity, visibility, and weather conditions to understand their impact on accident severity. The data is processed through a structured machine learning model that transforms these inputs into meaningful representations, capturing the relationships between situational factors and risk. By integrating this analysis with the initial driver-based evaluation, the model ensures a more comprehensive and accurate assessment, reflecting both individual behavior and external conditions.

In the price decision stage, the system combines the driver's risk profile with the predicted severity of potential accidents to calculate a proposed premium. Pricing is adjusted dynamically to align premiums with individual risk profiles, ensuring fairness for customers while optimizing profitability for insurers. This dynamic pricing mechanism rewards low-risk drivers with lower premiums and assigns higher premiums to high-risk profiles, creating an equitable and transparent pricing strategy.

To validate the system, an offline evaluation is conducted using synthetic datasets derived from real-world data. The dynamic pricing model is compared to a baseline fixed-rate model, assessing differences in accuracy, profitability, and alignment between premiums and actual risk levels. The results demonstrate the effectiveness of personalized pricing in reflecting individual risks and improving overall system performance.

3.1 Project Timeline

Our project was a semester long effort to get the required data and train the models to fulfill the requirement. Initially, the data provided by Bonzah included just Sales data and Claim Data of the accidents, however since this was not enough we ended up looking for public datasets to supplement our dataset. The US Accident US had the right information about location and weather but didn't have any information of the driver. So the first attempt was to try an embedding technique combining (See fig:3.2 all of the information together which didn't yield good results.

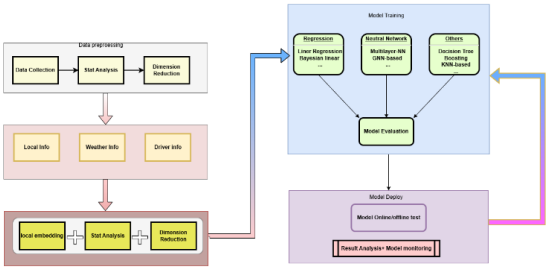

Figure 3.2 First Version

Since the first version didn't work out we developed a multi-task model, with embedding techniques to encode all the data(local, weather driver) separately.(See fig:3.3

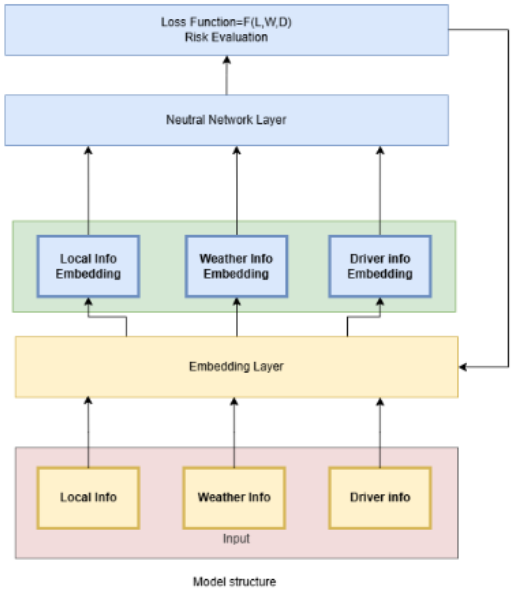

Figure 3.3 Seccond Version

Despite these improvements, the model didn't perform well,because of the limited driver information we had. The company then acquired additional driver data tailored to the customers of the company to supplement our dataset. After our research, we decided to use driver information as the main risk factor in accidents and the local and weather information as

secondary. This way we were able to develop a hierarchical model (See fig:3.9),in which driver information was processed separately and local and weather information separately.

3.2 Data Sources

Different datasets, including Location and Weather Information as well as Driver Information, are utilized, each serving a unique role in assessing driver risk. The Location and Weather data offer contextual factors that impact accident severity, while the Driver Information dataset focuses on individual behavior and history to provide a more accurate risk evaluation.

3.2.0.1 Location and Weather Information

This dataset was sourced from Kaggle US .It has over 7 million records from accident reports all over the USA. Given that the private rental insurance company operates nationwide, this dataset allows us to analyze the risk levels associated with different locations and conditions under which accidents occur. This enables us to identify high-risk and low-risk areas as well as specific conditions that contribute to accidents. This dataset includes over 40 columns, detailing the accident characteristics, environmental conditions, and geographic locations. However, due to the high dimensionality of the data and the processing constraints, only selected columns were used in our analysis. The feature importance chart (Figure 3.4) provides insights into which variables contribute most significantly to the model.We used a Random Forest Model to do the analysis. "Time Duration(min)," "Start Lat," and "Start Lng" are shown to be the top features, indicating that temporal and spatial factors are crucial in assessing accident severity. Environmental features like "Pressure(in)" and "Temperature(F)" also play an important role, further validating the impact of weather conditions on accident risks. This visualization guides the feature selection process and highlights the model's focus areas.

Initial analysis results are shown in Figure 3.5 and Figure 3.6, which display the states with the most severe accidents and the states with the highest number of accidents, respectively.

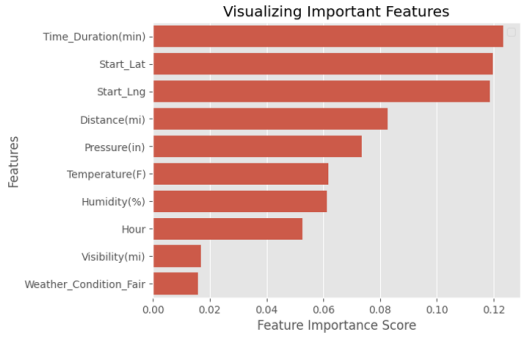

Figure 3.4 Feature Importance

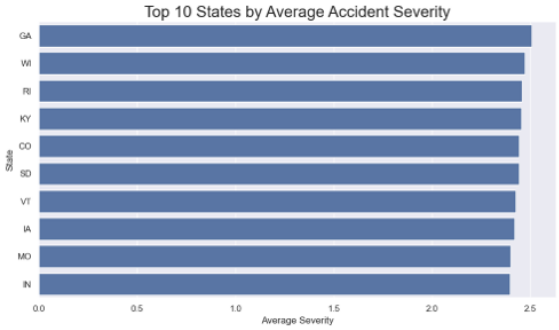

Figure 3.5 States with Most Severe Accidents

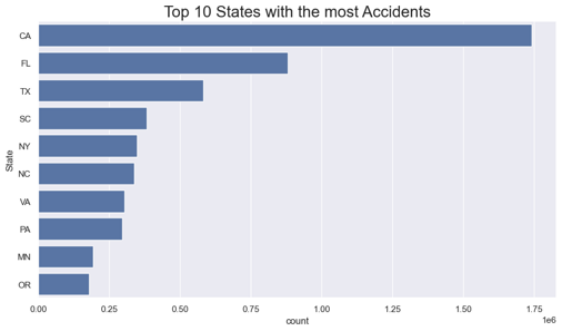

Figure 3.6 States with Most Accidents

3.2.0.2 Driver Information

The driver information dataset was gathered through the collaboration of the private insurance company - Bonzah with an external credit reporting company (TransUnion) .To construct a comprehensive profile for each driver, the company leveraged its sales data and sourced additional data related to each driver's violation history. This dataset includes several key attributes: Basic details such as the driver's state and age, A record of all major and minor traffic violations attributed to the driver, Specific instances where the driver was found guilty of a violation. including DUI violations and Whether a case related to a violation was dismissed or resulted in a guilty verdict.

The severity levels are as follows

- **Level 4**: Major violation count > 5.

- **Level 3**: $2 <$ Major violation count ≤ 5.
- **Level 2**: Major violation count ≥ 2 and < 5.
- **Level 1**: No major violations but more than 5 minor violations.
- **Level 0**: No major violations and less than 5 minor violations.

Because of data privacy concerns, data cannot be publically available, but the top features identified by a Random Forest Model are in the table. (See Table:3.1)

Feature Name	**Description**
NewlyAdjudicatedViolation	Insurance premium increased or not with newer violations.
TotalViolationSubRank1CountOtherState	Count of violations in other states with subrank1.
TotalViolationSubRank1MVRExcludedCount	Count of subrank1 violations excluded from Motor Vehicle Reports (MVR).
TotalMajorGuilty5yrCount	Total major violations with a guilty verdict in the last 5 years.

Table 3.1 Feature Names and Descriptions

Initial analysis on this dataset is in the figure 3.7 , 3.8

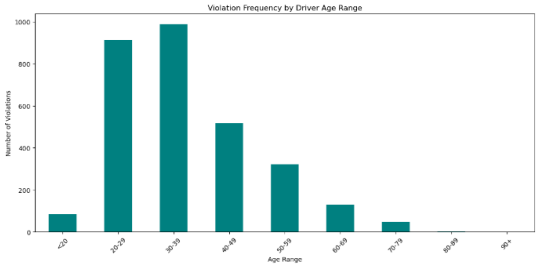

Figure 3.7 Violations by Age

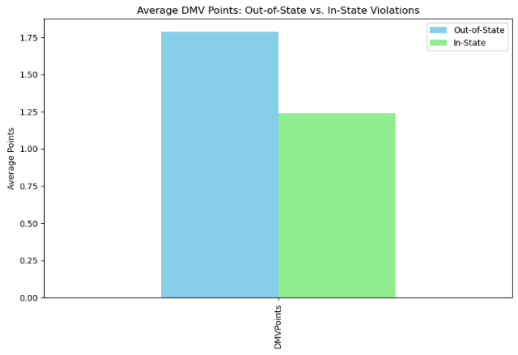

Figure 3.8 Violations by State

3.2.0.3 Sales Data

The Sales Data is a private dataset provided by the insurance company, containing records of all purchases made by customers. This dataset includes information on the types of insurance policies each customer opted for, along with details on coverage amounts, policy duration, and payment methods. This data helps us understand customer preferences, the range of coverage options selected, and patterns in purchasing behavior, which are essential for building predictive models on customer risk and pricing.This dataset is essentially merged with the driver data and local data.

3.3 Data Preprocessing

To prepare the datasets for modeling, we performed several key preprocessing steps. First, we converted all categorical columns to numerical representations, allowing the model to interpret categorical data effectively. We then removed rows containing any missing values to ensure data consistency and prevent issues during training due to the unavailability of the data .But in future we can use several data preprocessing techniques to ensure the missing data is properly handled ,such as replacing the values with the mean or median of the column. Finally, we applied scaling to the numerical columns, standardizing the feature ranges for improved model performance and interoperability.

SMOTE (Synthetic Minority Oversampling Technique) Chawla et al. (2002) was applied to the Location and Weather dataset due to the uneven distribution of data across different severity levels, which is a common issue in real-world datasets. The weather and location data often contain more instances of lower-severity cases and fewer high-severity events. This class imbalance leads to biased predictions and poor model performance, as the model tends to over-predict the majority class (e.g., low severity) and under-predict the minority class (e.g., high severity).

To address this imbalance, SMOTE was used to generate synthetic samples for the minority classes, effectively increasing the representation of the less frequent accident severities in the dataset. This allowed the model to learn more accurately from all classes, improving its ability to

predict both low- and high-severity accidents. By balancing the dataset, SMOTE helped ensure that the model could capture the full spectrum of possible outcomes, leading to better generalization and more accurate predictions for dynamic insurance pricing.

3.4 Models Explained

This project utilizes multiple models to assess the risk and determine the insurance premium for each driver. The primary models used are Neural Networks, Gradient Boosting Tree, XGBoost, which were selected due to their strong performance. These models process different datasets, including local and weather-related data and driver information, to predict accident risk and severity, which ultimately helps in calculating personalized insurance premiums.

3.4.1 Gradient Boosted Trees (GBT)

Gradient Boosting (GBT) is an ensemble learning technique that constructs a series of decision trees in a sequential manner, where each new tree tries to correct the errors made by the previous trees. The goal is to minimize the overall loss function, improving model accuracy with each iteration. This method is effective for both classification and regression tasks and can handle various types of data. However, GBT can be prone to overfitting, especially with too many trees or a high model complexity. Tuning hyperparameters such as the depth of the trees, the number of trees, and learning rates is crucial to achieving optimal performance and generalization.XGB (2020)

3.4.2 XGBoost

XGBoost (Extreme Gradient Boosting) is an advanced and optimized version of Gradient Boosting, designed to improve speed and efficiency. It enhances the standard gradient boosting algorithm by introducing regularization techniques to reduce overfitting and accelerate training with parallel processing. XGBoost is particularly well-suited for handling large datasets with complex structures and missing values. It has become popular in machine learning competitions

due to its high performance and scalability. By using XGBoost in the project, the goal is to leverage its ability to predict accident severity with high accuracy, and also predict driver risk making it a crucial part of the dynamic pricing model for insurance.XGB (2020)

3.4.3 Neural Networks

Neural Networks (NN) are used in this project to model complex, non-linear relationships between various features, such as weather, location, and driver information, to predict accident severity. By utilizing an embedding layer for categorical data and fully connected layers for processing, the neural network can capture intricate patterns in the data that simpler models might miss. NN excels in handling large datasets and learning hierarchical relationships, making it an ideal choice for this task.

On the left-side of the hierarchical model, (See fig:3.9), we have the accident prediction model that is trained on a combination of neural network and XGBoost. This is used to process the low-risk driver and act as a supplement to the driver risk identified. The weighted contribution is higher for the driver risk and lower for weather and location information.

For the driver risk assessment (See fig:3.9), two models - Neural Network and XGBoost were specifically trained to evaluate the risk associated with the driver profile, considering features like driving history, violations, and age. The structure of the neural network included ,input layer (driver information as input), Hidden Layers (with ReLU activations),Dropout Layer and Output Layer (predicting the accident severity)

Loss function used was Cross-Entropy and Optimiser Used was Adam Optimiser

3.5 Price Prediction

To estimate the expected insurance cost, our objective is to maximize $E[\text{price} \mid D, L, W]$, where D, L, and W represent driver, location, and weather information, respectively. Since the prior probability $P(D, L, W)$ is unknown, we assume a uniform distribution, $P(D, L, W) = c$, where c is a constant. This simplification allows us to express the expected value as:

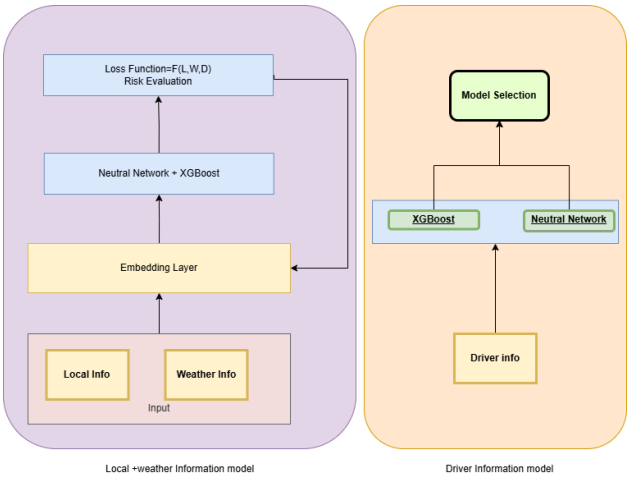

Figure 3.9 Models Used

$$
\begin{aligned}
E[\text{price} \mid D, L, W] &= \sum P(\text{price} \mid D, L, W) P(D, L, W) \\
&\sim P(\text{price} \mid D, L, W) \\
&= \sum P(\text{price} \mid S) P(S \mid D, L, W) \\
&\sim E(\text{price} \mid S) f(P(S \mid D), P(S \mid L, W))
\end{aligned}
$$

Here, S represents accident severity, ranging from 1 (minimal traffic impact) to 4 (significant disruption). The term $P(\text{price} \mid \text{severity})$ represents the likelihood of costs for each severity level. Our model focuses on predicting $P(\text{severity} \mid D, L, W)$ using public datasets, reducing the task to a classification problem for estimating insurance costs.

The function F is a linear function represented:

$$f(P(S \mid D), P(S \mid L, W)) = w_1 P(S \mid D) + w_2 P(S \mid L, W) \tag{3.1}$$

and here w1 and w2 are just set to 0.1 and 0.9

$$\text{subject to } \sum P(S \mid D) = 1, \sum P(S \mid L, W) = 1 \tag{3.2}$$

The multi-task model (See fig3.9 is used to develop the separate model needed to get the accurate pricing for the insurance premium.

3.6 Experimental Environment

We have used Python based machine learning frameworks : PyTorch, Cuda for Python 3.9 Nvidia 4060 Ti GPU was used to accelerate the training process.

3.7 Offline Evaluation

This project evaluates the performance of our price prediction model by comparing it to a baseline model. The baseline represents the current pricing policy used by Bonzah Insurance Company, which assigns a uniform price to all short-term renters regardless of individual characteristics.

The evaluation is conducted offline using simulated customer interactions, aiming to assess how effectively the models predict prices and maximize net income. Our framework (See fig:3.10) simulates real-world scenarios where customers provide their details, enabling models to predict prices based on input features. The primary evaluation metric, net income, is calculated as the difference between predicted prices and claim costs. The baseline model utilizes a static pricing strategy, often derived from historical averages (e.g., TransUnion data), offering simplicity and interpretability but lacking flexibility to adapt to dynamic factors such as driver profiles, location-specific conditions, and weather patterns. In contrast, our proposed model leverages machine learning to provide personalized pricing based on driver (D), location (L), and weather (W) information. This approach aims to adapt to dynamic conditions which means different prices for different drivers, locations and weather information, and improve pricing accuracy and profitability.

The offline evaluation simulates multiple customer interactions by randomly selecting customers from the dataset. For each interaction, the system mimics a customer inputting their details, including driver-related information, location, and prevailing weather conditions (retrieved via our API). Both the baseline model and the proposed machine learning model then generate price predictions. The claim cost for each interaction is predefined based on historical

accident and insurance data. Net income for each model is computed using the formula:

$$\text{Net Income} = \sum \text{Predicted Price} - \sum \text{Claim Cost}.$$

This metric evaluates the profitability of each model's pricing strategy. To ensure a comprehensive comparison, the evaluation will be the expectation of the net income across all simulated interactions.

In conclusion, the offline evaluation demonstrates the benefits of our machine learning-based price prediction model over traditional static pricing strategies. By utilizing data-driven methods, the proposed approach overcomes key limitations of conventional models. Future efforts will aim to tackle data sparsity issues and integrate additional factors, such as temporal trends and customer feedback, to improve the model's accuracy and practical relevance.

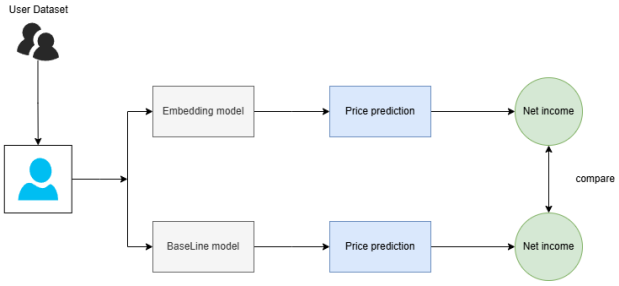

Figure 3.10 Offline Simulation

CHAPTER 4. RESULTS

4.1 Evaluation Metrics

We used Accuracy, Precision,, Recall and F1-Score as metrics for model comparison and analysis of the best model for the insurance pricing task.

The evaluation of the models demonstrated that the XGBoost model performed the best across the various datasets, achieving the highest accuracy and F1-score.4.1 The XGBoost model was particularly effective when predicting the price for local and weather-related information, significantly outperforming other models such as the neural network (NN) and gradient boosting models.

The driver-related model, particularly the XGBoost model, exhibited very high performance, with an accuracy of 99.76%, reflecting the model's ability to accurately predict driver risk factors based on historical violations.

Model/Task	**Accuracy**	**Precision**	**Recall**	**F1-Score**
WeatherNN (Separate)	41.41%	0.4071	0.4136	0.4056
LocationNN (Separate Learn)	41.50%	0.4194	0.4141	0.3735
Neural Network (Combined)	50.98%	0.5074	0.5098	0.4933
Embedding (Combined)	51.61%	0.5113	0.5162	0.5088
GradientBoost (Combined)	62.21%	0.6190	0.6221	0.6101
XGBoost (Combined)	71.61%	0.7102	0.7161	0.7029

Table 4.1 Location and Weather Model Performance Results

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Neural Network (NN)	98.66%	0.9877	0.9866	0.9869
XGBoost (XGB)	99.76%	0.9976	0.9976	0.9976

Table 4.2 Driver Model Performance

4.2 Offline Evaluation Results

The offline evaluation results show a comparison between the baseline method, the proposed method, and the optimal method across different simulations. The first graph 4.1 illustrates the performance of the baseline method versus the proposed method with an accident rate of 0.05. Here accident rate is an estimate made based on the sales data and the claim data. The baseline model demonstrates significant fluctuations in profits, whereas the proposed method stabilizes and provides more consistent results over time. This indicates that the dynamic pricing model proposed improves profitability by adjusting premiums based on risk predictions, compared to a flat-rate pricing structure used in the baseline.

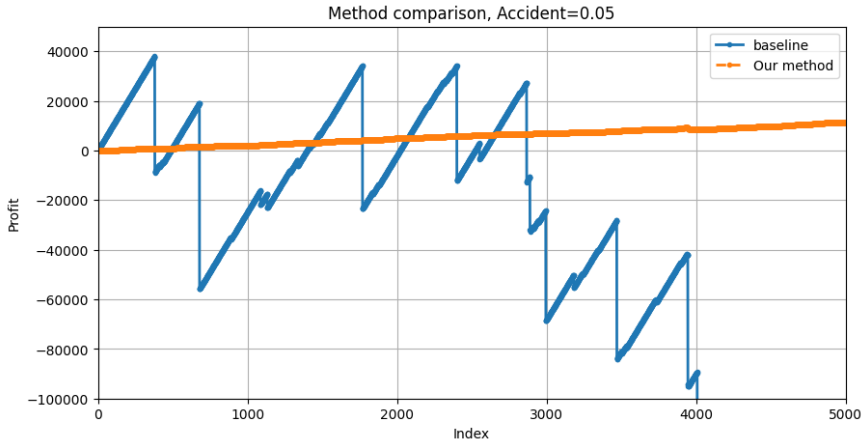

Figure 4.1 Method Comparison

The second graph 4.2 compares the results with the optimal values, showing that the proposed method's performance is closer to the optimal results than the baseline.(Optimal values refer to the theoretical or ideal performance that could be achieved if the model were able to perfectly adjust pricing based on the true risk and accident probabilities). The green line representing the optimal model stabilizes at a higher profit point, indicating that while the proposed method doesn't reach the optimal level, it still performs much better than the baseline model. This suggests that by dynamically adjusting premiums, the proposed method aligns closer to ideal profit levels, providing better financial outcomes for insurers than traditional methods.

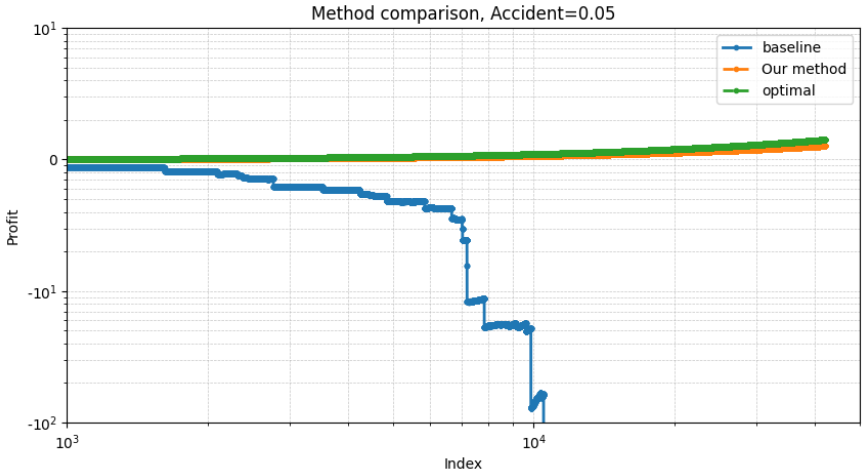

Figure 4.2 Method Comparison with Optimal Values

CHAPTER 5. SUMMARY AND DISCUSSION

This project focuses on the development and evaluation of a dynamic pricing model for car insurance, integrating multiple machine learning techniques to predict premiums based on driver risk and accident severity. The approach combines a variety of models, including neural networks, XGBoost, and gradient boosting, which utilize both local/weather information and driver-specific data. The models were trained to predict the likelihood of an accident and its potential severity, to adjust the insurance price accordingly. The driver data, such as age, violations, and history, and the local data, including weather and geographical factors, were essential in refining the model's predictions.

In the evaluation phase, the dynamic pricing method performed significantly better than the baseline pricing approach which used a fixed price for all the driver profiles, regardless of the risk. This makes the premiums that were adjusted using our approach more profitable pricing strategies. Our personalized pricing model came significantly closer to the optimal pricing value and offers a clear advantage over the traditional methods. These findings suggest that dynamic, data-driven pricing can lead to more efficient and fair pricing, benefitting both insurers and policyholders.

5.1 Future Work

In future work, we aim to build on these findings by:

1. **Enhanced Data Collection**: We aim to gather more comprehensive datasets to better represent the complex interactions among features. Currently, our dataset is limited to 20,000 sales records and 200 claim records, supplemented with public datasets for adjustments. Expanding data sources will strengthen the robustness of our predictions.

2. **Model Refinement**: We intend to improve the multi-task architecture by incorporating advanced deep learning techniques such as attention mechanisms and graph-based learning. While current performance improvements remain marginal, these methods can enhance feature engineering and data augmentation, providing greater flexibility for future enhancements.

3. **Real-Time Deployment**: Our goal is to develop a dynamic pricing engine capable of real-time predictions and adjustments based on evolving customer inputs and environmental conditions. This extension will enable the model to function as an online learning system, adapting to new information continuously.

4. **Fairness and Bias Audits**: We will prioritize fairness analysis to address potential biases across demographic and geographic segments. Ensuring equitable predictions is critical for building trust and compliance in commercial applications.

BIBLIOGRAPHY

Us accidents (2016 - 2023).

(2020). Gbt and xgboost. Available at: https://medium.com/@gabrieltseng/gradient-boosting-and-xgboost-c306c1bcfaf5.

(2024). Safe driving discounts. Available at: https://www.cnbc.com/select/safe-driving-discounts-how-do-they-work/.

Alamir, E., Urgessa, T., Hunegnaw, A., and Gopikrishna, T. (2021). Motor insurance claim status prediction using machine learning techniques. *International Journal of Advanced Computer Science and Applications*, 12.

Ayuso, M., Guillén, M., and Marín, A. M. P. (2016). Using gps data to analyse the distance travelled to the first accident at fault in pay-as-you-drive insurance. *Transportation Research Part C: Emerging Technologies*, 68.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Hanafy, M. and Ming, R. (2021). Improving imbalanced data classification in auto insurance by the data level approaches. *International Journal of Advanced Computer Science and Applications*, 12.

Pesantez-Narvaez, J., Guillen, M., and Alcañiz, M. (2019). Predicting motor insurance claims using telematics data—xgboost versus logistic regression. *Risks*, 7.

Power, J., Côté, M. P., and Duchesne, T. (2024). A flexible hierarchical insurance claims model with gradient boosting and copulas. *North American Actuarial Journal*.

Verbelen, R., Antonio, K., and Claeskens, G. (2018). Unravelling the predictive power of telematics data in car insurance pricing. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 67.

Wang, H. D. (2020). Research on the features of car insurance data based on machine learning. In *Procedia Computer Science*, volume 166.

Xie, S. (2021). Improving explainability of major risk factors in artificial neural networks for auto insurance rate regulation. *Risks*, 9.