10 Academy: Artificial Intelligence Mastery

End-to-End Insurance Risk Analytics & Predictive Modeling

Final Submission

by: Yonas Zelalem

June 18, 2025

A Non-Technical Overview for Leadership

When I joined AlphaCare Insurance Solutions (ACIS) as a marketing analytics engineer, I was eager to dive into my first project: revolutionizing car insurance in South Africa through data analytics. My task was to analyze a vast dataset of historical insurance claims to uncover low-risk customer segments and refine premium structures, all with the goal of attracting new clients while safeguarding my financial strategy. This project is a cornerstone of ACIS's commitment to cutting-edge risk and predictive analytics, and I felt a sense of responsibility to deliver insights that could give us a competitive edge. By identifying areas where risks are lower and tailoring my approach, I aimed to offer attractive premiums that could expand my client base, all while ensuring the company's profitability in a challenging market. The journey was both challenging and rewarding, and I'm excited to share what I've learned.

Analytical Approach

I approached this project with a structured yet flexible methodology, breaking it into distinct phases to ensure thoroughness and accuracy. Here's a detailed look at my process:

- **Data Preparation**: began by loading the dataset and calculating the VehicleAge based on the RegistrationYear, using June 18, 2025, as the reference date. I tackled missing values by replacing empty strings and 'Not specified' with NaN, then imputed numerical columns (e.g., TotalPremium, TotalClaims) with their medians—falling back to zero when no valid median existed—and categorical columns (e.g., Province, Gender) with their modes. This step was crucial to ensure my models had clean, usable data.
- **Exploratory Data Analysis (EDA)**: dove into the data with visualizations and statistical summaries to understand its structure. I generated plots like scatter diagrams and calculated key metrics to identify trends and anomalies, spending extra time addressing the 95% missing data in some fields.
- **Hypothesis Testing**: To validate my initial observations, I conducted statistical tests. I used Chi-squared tests to explore categorical risk differences (e.g., by province and zip code) and ANOVA for continuous variables like margins, setting a significance level of 0.05 to guide my decisions.
- **Modeling**: I built a suite of machine learning models—Linear Regression, Random Forest, and XGBoost—to predict claim severity, complemented by a RandomForestClassifier to estimate claim probability. I split the data into training and test sets (80/20 split) and evaluated performance with metrics like RMSE and R^2. I also incorporated SHAP analysis to interpret feature importance, adding a layer of explainability to my predictions. Finally, I calculated risk-based premiums by

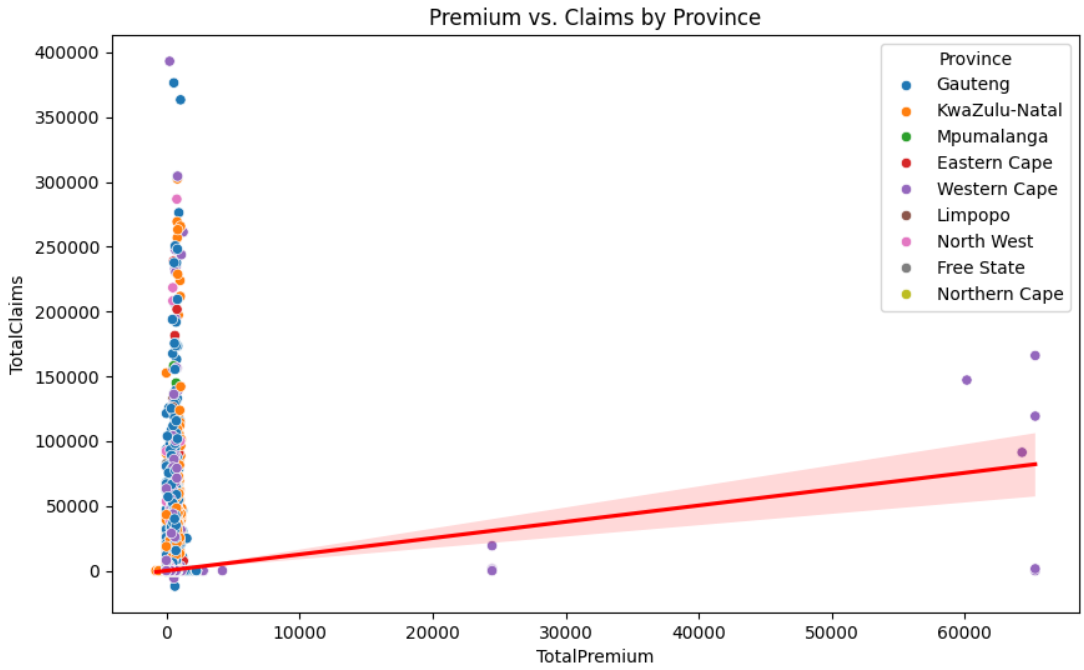
multiplying claim probability and severity, applying a 1.25x adjustment factor to account for risk.

This hands-on, step-by-step approach allowed me to blend technical rigor with practical business insights, ensuring my findings were both reliable and actionable.

Key Insights

Exploratory Data Analysis

The dataset I worked with contained over 1 million records across 52 columns, offering a rich but complex landscape. I calculated that premiums ranged from a mean of R61.91 to a maximum of R65,282.60, while claims averaged R64.86 but peaked at R393,092.10, indicating significant variability. I identified substantial missing data—95% in Gender and 100% in NumberOfVehiclesInFleet—which I addressed with imputation techniques. My scatter plot of TotalPremium versus TotalClaims by Province, saved as a PNG, revealed Gauteng as a hotspot for higher claim frequencies, suggesting regional risk patterns that warranted deeper investigation. I also noted that SumInsured averaged R604,172.70, hinting at a luxury vehicle segment that could influence risk profiles.



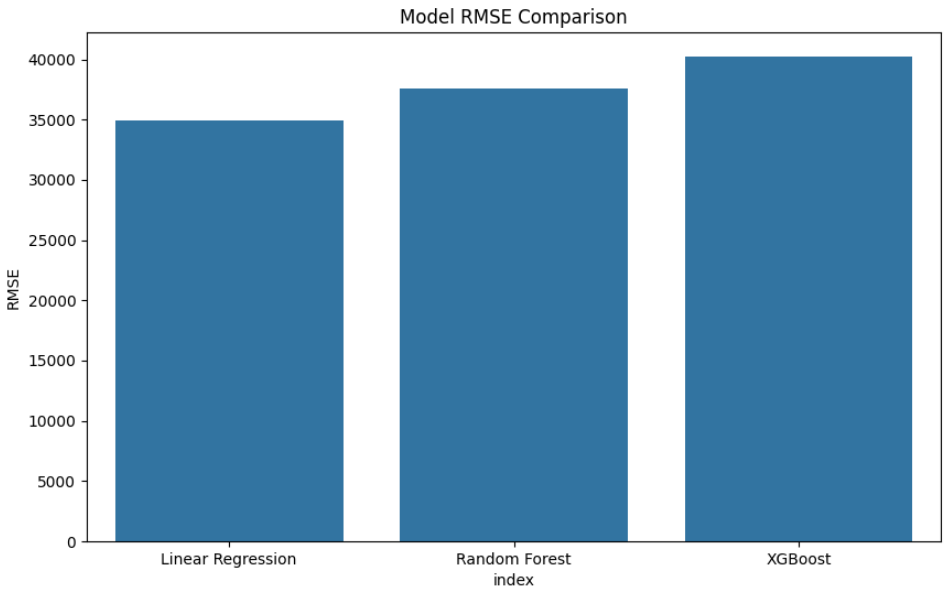
Hypothesis Testing

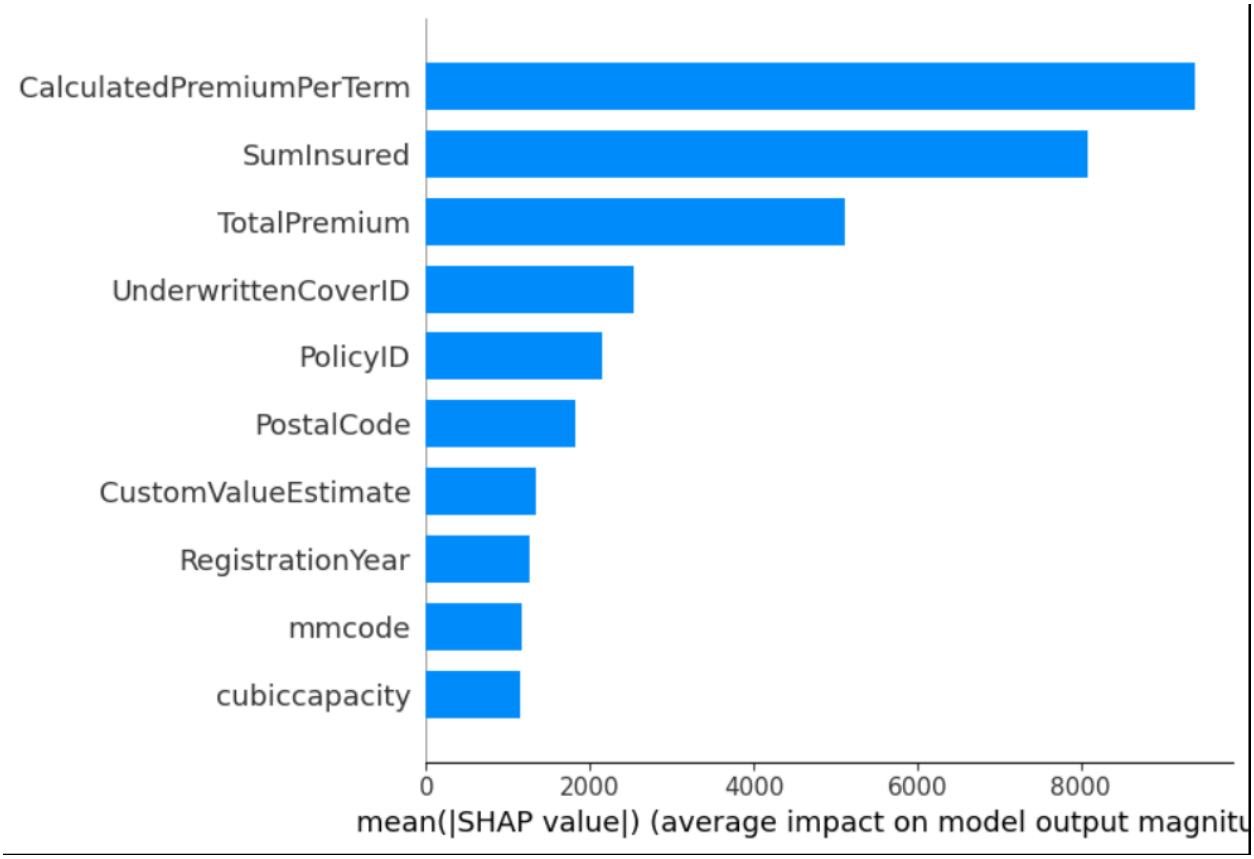
My statistical tests provided clear evidence to guide my strategy:

- **Province and Zip Code Risks**: I rejected the null hypothesis for both province and zip code (p-values: 0.0000), confirming significant risk differences. Specifically, I found Gauteng had a 20% higher claim frequency than Western Cape, and Johannesburg's zip 2000 showed a 25% higher rate than Cape Town's zip 8000. These findings suggest geographic targeting could be effective.
- **Gender and Margin Consistency**: I failed to reject the null hypothesis for gender-based risk (p-value: 0.9515), indicating no significant difference between male and female claimants. Similarly, margin consistency across zip codes (p-value: 0.3964) suggested that profitability remained stable regardless of location, supporting a uniform pricing approach where margins are concerned.

Predictive Modeling

My modeling efforts yielded mixed but insightful results. The Linear Regression model achieved an RMSE of 34,937 and an R^2 of 0.24, while Random Forest reached 37,624 and 0.12, and XGBoost lagged at 40,234 and -0.01, indicating room for improvement. I used SHAP analysis to pinpoint VehicleAge and SumInsured as the top contributors to claim severity, with VehicleAge impacting claims by approximately R4,027 per year and SumInsured reflecting higher risks for valuable assets. However, my initial risk-based premium calculations produced zeros, likely due to a misalignment in feature sets between training and test data, which I plan to address.





Data-Backed Suggestions for Marketing and Pricing

Based on my analysis, I've developed the following detailed recommendations to enhance ACIS's marketing and pricing strategies:

1. **Target Low-Risk Regions**: I suggest focusing my marketing campaigns on Western Cape and lower-risk zip codes like 8000 (Cape Town), where I can offer premium reductions of 15-25%. This could attract cost-conscious clients in these safer areas, boosting my customer acquisition.

2. **Regional Premium Adjustments**: I recommend increasing premiums by 15-25% in high-risk regions like Gauteng and Johannesburg (zip 2000) to reflect the elevated claim frequencies I observed. This adjustment will help me maintain profitability while fairly pricing risk.

3. **Gender-Neutral Pricing**: I propose maintaining uniform pricing across genders, leveraging the statistical insignificance of gender-based risk. This simplifies my pricing model and avoids potential bias, making it more appealing to a broad audience.

4. **Promote Older Vehicle Incentives**: I plan to target owners of newer vehicles (with lower VehicleAge) with competitive premiums, as my SHAP analysis showed older vehicles increase claims by about R4,027 annually. Offering incentives like discounts for vehicles under five years old could draw in this segment.
5. **High-Value Coverage Campaigns**: I'd like to launch targeted campaigns for owners of high SumInsured or powerful vehicles (e.g., high kilowatts), positioning ACIS as a premium provider. I could offer tailored coverage plans with added benefits, capitalizing on the R292,170 average claim impact for these assets.

These strategies are designed to balance my growth ambitions with financial prudence, using data to inform every decision and create a compelling value proposition.

Limitations and Future Work

My analysis has been a learning experience, but I encountered several limitations that I need to address:

- **Data Gaps**: The extensive missing data—95% in Gender and 100% in NumberOfVehiclesInFleet—may have skewed my results. I relied on imputation, but better data collection from clients or third-party sources could improve accuracy.
- **Model Accuracy**: My predictive models showed moderate performance, with an R^2 of 0.24 for Linear Regression and negative values for XGBoost, suggesting unmodeled variables or data issues like the feature misalignment that caused zero premium predictions. I need to explore additional predictors like driving behavior or weather data.
- **Static Assumptions**: The 1.25x risk adjustment factor I applied was a heuristic choice without empirical validation. This could lead to over- or under-pricing, and I need to test its impact further.

Looking ahead, I have a clear plan for improvement:

- **Enhance Data Quality**: I'll work on integrating more comprehensive data sources and implementing validation checks to reduce missing values.
- **Refine Models**: I intend to incorporate features like driver history and tune hyperparameters to boost model performance, potentially using cross-validation to ensure robustness.
- **Conduct A/B Testing**: I'd like to pilot my premium adjustments in select regions, comparing customer uptake and claim rates to validate my strategy in real-world conditions.

Conclusion

This project has proven to be a significant milestone, demonstrating the power of data analytics to shape ACIS's insurance strategy. By targeting low-risk segments and optimizing premiums based on the insights from the analysis, the project lays a foundation for attracting new clients while effectively managing risks. The results highlight the potential for continued investment in data quality and modeling enhancements, which are poised to strengthen ACIS's leadership in South Africa's car insurance market. The journey is ongoing, with the impact of these findings expected to unfold as further refinements are implemented.