Interim Report

Insurance Risk Analytics & Predictive Modeling

AlphaCare Insurance Solutions (ACIS)

Report Date: December 06, 2025

Covering Task 1 (EDA) and Task 2 (DVC Setup)

Table of Contents

1. Executive Summary

This interim report presents the progress made on the Insurance Risk Analytics and Predictive Modeling project for AlphaCare Insurance Solutions (ACIS). The project aims to analyze historical insurance claim data to identify low-risk segments and optimize premium pricing strategies. This report covers two completed tasks: (1) comprehensive Exploratory Data Analysis (EDA) of insurance data spanning February 2014 to August 2015, and (2) establishment of a Data Version Control (DVC) system for reproducible and auditable data pipelines. Key highlights include the identification of significant loss ratio variations across provinces, vehicle types, and demographic segments. The overall portfolio loss ratio of 104.77% indicates that claims exceed premiums, highlighting the critical need for risk-based premium optimization. The DVC infrastructure ensures regulatory compliance and reproducibility, essential for financial services analytics.

2. Understanding and Defining the Business Objective

2.1 Business Context

AlphaCare Insurance Solutions operates in the competitive South African car insurance market. The company's strategic objective is to develop data-driven risk analytics capabilities that enable more precise premium pricing and targeted marketing strategies. In an industry where profitability depends on accurately assessing and pricing risk, the ability to identify low-risk customer segments presents a significant competitive advantage.

2.2 Primary Objectives

- Discover 'low-risk' target segments for premium reduction opportunities
- Build predictive models for optimal premium pricing based on risk factors
- Optimize marketing strategies by targeting profitable customer segments
- Ensure regulatory compliance through auditable and reproducible analytics

2.3 Success Metrics

- Loss Ratio (TotalClaims / TotalPremium) by segment
- Risk differentiation across geographic regions (Provinces, Postal Codes)
- Risk patterns by vehicle characteristics (Type, Make, Model)
- Demographic risk factors (Gender, Marital Status)
- Temporal trends in claims frequency and severity

3. Task 1: Exploratory Data Analysis (EDA)

3.1 Data Overview

The analysis utilized historical insurance data from February 2014 to August 2015, containing 1,000,098 transaction records across 52 original features. The dataset includes comprehensive information about policies, clients, vehicle characteristics, geographic locations, and financial metrics (premiums and claims).

3.2 Data Quality Assessment

Data quality assessment revealed several important characteristics:

- Missing values identified in 20+ columns, with highest missing rates in fleet-related fields (100%)
- CustomValueEstimate missing in 77.96% of records, requiring careful handling in modeling
- No duplicate rows detected, indicating clean data ingestion
- Outliers detected in TotalClaims (0.28%) and CustomValueEstimate (0.81%) using IQR method

3.3 Key Financial Metrics

Metric	Value
Total Premium	ZAR 61,911,562.70
Total Claims	ZAR 64,867,546.17
Overall Loss Ratio	104.77%
Data Period	Feb 2014 - Aug 2015
Total Records	1,000,098

3.4 Loss Ratio Analysis by Segment

3.4.1 By Province

Significant variation in loss ratios across provinces was identified:

Province	Loss Ratio	Risk Level
Gauteng	122.20%	High Risk
KwaZulu-Natal	108.27%	High Risk
Western Cape	105.95%	High Risk

North West	79.04%	Moderate Risk
Mpumalanga	72.09%	Low Risk
Free State	68.08%	Low Risk
Limpopo	66.12%	Low Risk
Eastern Cape	63.38%	Low Risk
Northern Cape	28.27%	Very Low Risk

3.4.2 By Vehicle Type

Vehicle type analysis reveals distinct risk profiles:

- Heavy Commercial vehicles show highest loss ratio (162.81%) - critical risk segment
- Passenger Vehicles and Medium Commercial both exceed 100% loss ratio
- Light Commercial (23.21%) and Bus (13.73%) represent low-risk opportunities

3.4.3 By Gender

Gender-based analysis indicates:

- Not specified category shows highest loss ratio (105.93%)
- Male drivers: 88.39% loss ratio
- Female drivers: 82.19% loss ratio - lowest risk segment

3.5 Temporal Trends

Analysis of monthly trends from February 2014 to August 2015 reveals:

- Loss ratio peaked in April 2015 at 139.64%
- Significant volatility in claims frequency over the 18-month period
- Premium collection remained relatively stable
- August 2015 shows unusually low loss ratio (14.01%) - requires investigation

3.6 Vehicle Make Analysis

Top vehicle makes by total claims reveal significant risk concentration:

- TOYOTA: Highest total claims (ZAR 51.7M) with 103.60% loss ratio
- AUDI: 271.35% loss ratio - extremely high risk

- HYUNDAI: 398.98% loss ratio - critical risk segment
- MERCEDES-BENZ: 106.29% loss ratio with ZAR 2.9M in claims

4. Task 2: Data Version Control (DVC) Setup

4.1 Objective

Established a reproducible and auditable data pipeline using Data Version Control (DVC) to meet regulatory compliance requirements in the financial services industry. DVC ensures that data inputs are as rigorously version-controlled as code, enabling complete reproducibility of analyses and models for auditing, regulatory compliance, and debugging purposes.

4.2 Implementation

The DVC infrastructure was successfully implemented with the following components:

- DVC repository initialized in project directory
- Local remote storage configured at './dvc_storage/'
- Primary data file (MachineLearningRating_v3.txt, ~529 MB) added to DVC tracking
- Data file metadata (.dvc file) committed to Git repository
- Actual data file stored in DVC remote storage, excluded from Git
- Configuration files (.dvc/config, .dvcignore) properly version-controlled

4.3 Benefits

- Reproducibility: Any analysis can be reproduced using exact data versions
- Auditability: Complete history of data changes for regulatory compliance
- Storage Efficiency: Large files stored outside Git repository
- Version Control: Track different versions of datasets over time
- Collaboration: Team members can pull specific data versions as needed

5. Key Findings and Insights

5.1 Critical Business Insights

- Overall portfolio is unprofitable with 104.77% loss ratio - immediate action required
- Geographic segmentation reveals 3x risk difference between highest (Gauteng: 122.20%) and lowest (Northern Cape: 28.27%) risk provinces
- Vehicle type segmentation shows 12x risk difference (Heavy Commercial: 162.81% vs Bus: 13.73%)
- Female drivers represent lower risk segment (82.19% vs 88.39% for males)
- Specific vehicle makes (AUDI, HYUNDAI) show extreme risk profiles requiring targeted pricing

5.2 Low-Risk Opportunities

- Northern Cape, Eastern Cape, Limpopo provinces - premium reduction opportunities
- Bus and Light Commercial vehicle types - expand market share
- Female driver segment - targeted marketing campaigns
- Specific vehicle makes with low loss ratios - competitive pricing strategies

5.3 High-Risk Segments Requiring Action

- Gauteng, KwaZulu-Natal, Western Cape provinces - premium adjustments needed
- Heavy Commercial vehicles - risk-based pricing implementation
- AUDI and HYUNDAI vehicle makes - underwriting review required
- Not specified gender category - data quality improvement needed

6. Next Steps and Key Areas of Focus

6.1 Immediate Next Steps

- A/B Hypothesis Testing: Validate risk differences across provinces, zipcodes, and gender
- Statistical Modeling: Develop linear regression models per zipcode to predict total claims
- Machine Learning Pipeline: Build predictive model for optimal premium pricing
- Feature Engineering: Create derived features from existing variables
- Model Evaluation: Assess model performance and feature importance

6.2 Hypothesis Testing Priorities

- H0: No risk differences across provinces → Reject based on EDA findings
- H0: No risk differences between zipcodes → Requires statistical validation
- H0: No margin (profit) differences between zip codes → Critical for pricing strategy
- H0: No significant risk difference between Women and Men → Preliminary evidence suggests difference exists

6.3 Modeling Priorities

- Linear Regression per Zipcode: Predict TotalClaims using local risk factors
- Premium Optimization Model: ML model incorporating car features, owner demographics, location, and other relevant factors
- Feature Importance Analysis: Identify key drivers of risk and profitability
- Model Interpretability: Ensure models are explainable for regulatory compliance

6.4 Data Pipeline Enhancements

- Implement data validation checks in DVC pipeline
- Create automated EDA reports for new data versions
- Set up data quality monitoring dashboards
- Establish data versioning best practices documentation

7. Conclusion

This interim report demonstrates significant progress in establishing the foundation for data-driven risk analytics at AlphaCare Insurance Solutions. The comprehensive EDA has revealed critical insights into risk segmentation across geographic, vehicle, and demographic dimensions. The identification of low-risk segments (Northern Cape, Bus vehicles, Female drivers) presents immediate opportunities for premium optimization and market expansion. The establishment of DVC infrastructure ensures that all future analyses will be reproducible and auditable, meeting the stringent requirements of financial services regulation. This foundation enables confident progression to hypothesis testing and predictive modeling phases. The findings from this analysis provide a clear roadmap for risk-based pricing strategies and targeted marketing initiatives. The next phase of work will focus on statistical validation of these insights and development of predictive models for premium optimization. The project is on track to deliver actionable recommendations that will enable ACIS to improve profitability through data-driven risk assessment and premium pricing optimization.

Appendix

Detailed visualizations and analysis outputs are available in the outputs/figures/ directory. Key visualizations include loss ratio heatmaps, risk-return scatter plots, temporal evolution dashboards, and distribution analyses. All code and analysis notebooks are version-controlled in the project repository.