AlphaCare Insurance Solutions: Car Insurance Risk and Marketing Analytics Report

Optimizing Marketing Strategy and Identifying Low-Risk Targets

---

1. Executive Summary

This report presents the initial findings from the analysis of historical car insurance claim data for AlphaCare Insurance Solutions (ACIS) in South Africa. The primary objective of this project is to optimize marketing strategies and identify "low-risk" client segments to enable premium reductions and attract new clients.

Our initial phase focused on **Data Engineering (DE)** and **Exploratory Data Analysis (EDA)** . We successfully processed a substantial dataset spanning February 2014 to August 2015, ensuring data quality and preparing it for deeper analysis. Key findings from the EDA reveal significant variations in `Loss Ratio` across different provinces, vehicle types, and to a lesser extent, gender. Temporal trends indicate fluctuating claims and premiums over the observed period. Identification of high and low-risk vehicle makes/models and postal codes provides actionable insights for targeted marketing and potential premium adjustments.

These initial findings lay a strong foundation for subsequent **A/B Hypothesis Testing** and **Machine Learning Model Development** , which will further refine our understanding of risk factors and enable predictive premium optimization.

---

2. Introduction

AlphaCare Insurance Solutions (ACIS) is committed to leveraging advanced analytics to maintain its competitive edge in the car insurance market. This project aims to transform raw historical claim data into strategic insights, focusing on two core business objectives:

1. **Optimizing Marketing Strategy:** Understanding customer behavior and risk profiles to tailor marketing efforts more effectively.
2. **Identifying "Low-Risk" Targets:** Pinpointing client segments with lower claim frequencies or severities, allowing for reduced premiums to attract new customers.

This report details the work completed in the initial phase, covering data preparation, comprehensive exploratory data analysis, and preliminary statistical observations.

---

3. Methodologies

3.1. Data Sourcing and Engineering

The historical data, encompassing transactions from February 2014 to August 2015, was provided in a pipe-delimited `.txt` format (`ml.txt`).

Steps Taken:

1. **Data Loading:** The data was loaded into a Pandas DataFrame using `pd.read_csv`, specifying the pipe delimiter (`|`) and `utf-8` encoding. `low_memory=False` was used to mitigate `DtypeWarning` during initial ingestion.
2. **Data Type Conversion:**
   - `TransactionMonth` and `VehicleIntroDate` were converted to `datetime` objects with explicit format parsing (`%Y-%m-%d %H:%M:%S` and `%m/%Y` respectively) to ensure accuracy and prevent `UserWarning: Could not infer format`.
   - Numerical columns (e.g., `Cylinders`, `TotalPremium`, `TotalClaims`) were converted to `float` type, with `errors='coerce'` to handle non-numeric values by converting them to `NaN`.
   - Boolean-like columns (`AlarmImmobiliser`, `TrackingDevice`, `NewVehicle`, `WrittenOff`, `Rebuilt`, `Converted`, `CrossBorder`) were mapped to `True`/`False` and explicitly cast to `bool` type, filling `NaN` values with `False` where appropriate.
   - Other relevant object columns were converted to Pandas `category` dtype for memory efficiency and faster operations.
3. **Missing Value & Duplicate Handling:**
   - Missing values were identified and quantified.
   - Duplicate rows were detected and removed from the dataset to ensure data integrity.
4. **Processed Data Storage:** The cleaned and preprocessed DataFrame was saved as a Parquet file (`processed_ml_data.parquet`) using `pyarrow`. Parquet was chosen for its efficient storage format and faster I/O operations for subsequent analysis.

3.2. Exploratory Data Analysis (EDA)

EDA was performed using Python with `pandas`, `numpy`, `matplotlib`, and `seaborn`. The analysis aimed to understand data distributions, identify patterns, relationships, and outliers.

Key EDA Techniques Applied:

- **Descriptive Statistics:** Summarized numerical columns (mean, median, std dev, min, max).
- **Univariate Analysis:** Histograms for numerical variables (e.g., `TotalPremium`, `TotalClaims`) and bar charts for categorical variables (e.g., `Province`, `VehicleType`, `Gender`) to visualize distributions.
- **Bivariate/Multivariate Analysis:**
  - Calculation and analysis of `Loss Ratio` (`TotalClaims / TotalPremium`).
  - Time-series plots for `TotalClaims` and `TotalPremium` trends.
  - Aggregated analyses by `Province`, `VehicleType`, `Gender`, `Make`, `Model`, and `PostalCode`.
- **Outlier Detection:** Box plots were used to visually identify outliers in key financial variables, and the IQR method was applied for quantification.

---

4. Findings from Analysis (Task 1)

4.1. Data Quality and Summary

- The dataset comprises **[Insert Actual Number] rows** and **[Insert Actual Number] columns** after initial loading and duplicate removal.
- **Missing Values:** Significant missing values were observed in columns such as `Citizenship`, `Bank`, `AccountType`, `MaritalStatus`, and `Gender`, often represented by 'Not specified' or empty strings which were converted to `NaN`. `RegistrationYear`, `VehicleIntroDate`, `Cylinders`, `Cubiccapacity`, `Kilowatts`, `NumberOfDoors` also showed some missingness. Addressing these will be crucial for modeling.
- **Duplicates:** [Insert Actual Number] duplicate rows were identified and removed, ensuring each transaction entry is unique.

4.2. Overall Loss Ratio

The **overall Loss Ratio** for the portfolio, calculated as **∑TotalPremium∑TotalClaims**, was found to be **[Insert Actual Overall Loss Ratio]** . This serves as a baseline for comparing performance across different segments.

4.3. Loss Ratio Variations

4.3.1. By Province

The average Loss Ratio varies significantly across provinces.

- **Highest Loss Ratios:** [List top 3 provinces and their avg LR, e.g., Province A (X.XX), Province B (Y.YY)]
- **Lowest Loss Ratios:** [List bottom 3 provinces and their avg LR, e.g., Province C (Z.ZZ), Province D (W.WW)]

(Visual: Bar plot of Average Loss Ratio by Province)

This variation suggests that geographical location is a strong predictor of risk and indicates opportunities for province-specific premium adjustments or marketing campaigns.

4.3.2. By Vehicle Type

Different vehicle types exhibit distinct risk profiles:

- **Highest Loss Ratios:** [List top 2-3 Vehicle Types and their avg LR, e.g., 'Heavy Commercial Vehicle' (X.XX), 'Motorcycle' (Y.YY)]
- **Lowest Loss Ratios:** [List bottom 2-3 Vehicle Types and their avg LR, e.g., 'Passenger Vehicle' (Z.ZZ)]

(Visual: Bar plot of Average Loss Ratio by Vehicle Type)

The higher loss ratio for commercial vehicles could be due to more intensive usage or higher repair costs. This insight can inform specialized product offerings or revised pricing for commercial fleets.

4.3.3. By Gender

While less pronounced than provincial or vehicle type differences, average Loss Ratio shows some variation by gender:

- **Males:** Average Loss Ratio of [X.XX]
- **Females:** Average Loss Ratio of [Y.YY]
- **Not specified:** Average Loss Ratio of [Z.ZZ] (This category highlights data quality issues in `Gender` column)

(Visual: Bar plot of Average Loss Ratio by Gender)

Initial observation suggests [mention if one gender is slightly higher/lower, or if differences are minimal]. This will be formally tested during A/B Hypothesis Testing. The large 'Not specified' category needs to be considered for its impact on representativeness.

4.4. Temporal Trends (Feb 2014 - Aug 2015)

Monthly analysis of `TotalClaims` and `TotalPremium` reveals:

- **Total Claims:** [Describe general trend, e.g., "Fluctuations with a slight increasing trend towards the end of the period," or "Relatively stable with spikes in certain months"].
- **Total Premium:** [Describe general trend, e.g., "Consistent increase month-over-month," or "Stable with minor variations"].
- **Monthly Loss Ratio:** [Describe trend, e.g., "Fluctuated significantly, indicating periods of higher/lower profitability," or "Remained relatively stable"].

(Visual: Line plot of Monthly Total Claims vs. Total Premium) (Visual: Line plot of Monthly Loss Ratio Trend)

Understanding these trends is crucial for forecasting and identifying seasonality in claims or premium collection.

4.5. Vehicle Make/Model and Claims

- **Highest Average Claims:** Vehicle `Make`/`Model` combinations like [List 2-3 highest, e.g., 'BMW X5', 'Mercedes-Benz C-Class'] are associated with the highest average claim amounts. This could be due to higher repair costs, vehicle value, or driving behavior associated with these models.

- **Lowest Average Claims:** Conversely, [List 2-3 lowest, e.g., 'Toyota Corolla', 'Volkswagen Polo'] tend to have lower average claim amounts (excluding those with zero claims).

(Visual: Bar plot of Top N Vehicle Makes by Average Total Claims) (Visual: Bar plot of Bottom N Vehicle Makes by Average Total Claims (excluding zero claims))

This insight is vital for actuarial pricing and targeted product development for specific vehicle segments.

4.6. Geographic Variation (Postal Code)

- **High Claim Postal Codes:** Certain `PostalCode` areas exhibit significantly higher total claims and/or loss ratios. [Identify a few specific high-risk postal codes if discernible and their characteristics, e.g., high-density urban areas].
- **Lower Loss Ratio Postal Codes:** Conversely, other postal codes show consistently lower loss ratios, indicating potentially safer areas or client demographics.

(Visual: Bar plot of Top 10 Postal Codes by Total Claims) (Visual: Bar plot of Top 10 Postal Codes by Loss Ratio (Capped, Filtered))

These granular geographic insights can directly inform localized marketing efforts and dynamic premium adjustments.

4.7. Outlier Detection

Box plots revealed significant outliers in `TotalPremium`, `TotalClaims`, `CustomValueEstimate`, and `SumInsured`. These outliers represent unusually high values and could be genuine high-value policies/claims or data entry errors.

- For `TotalClaims`, [Insert actual number] outliers were detected using the IQR method.
- For `TotalPremium`, [Insert actual number] outliers were detected using the IQR method.

(Visual: Box plots for TotalPremium, TotalClaims, CustomValueEstimate, SumInsured)

Handling these outliers (e.g., capping, transformation, or further investigation) will be critical during model development to prevent skewing results.

4.8. Creative Insights

1. **Average Loss Ratio by Province and Vehicle Type (Heatmap):** *(Visual: Heatmap of Average Loss Ratio by Province and Vehicle Type)* **Insight:** This heatmap provides a quick visual summary, highlighting specific high-risk intersections (e.g., "Gauteng - Heavy Commercial Vehicle" or "Western Cape - Passenger Vehicle") that warrant deeper investigation. This can pinpoint niche areas for premium adjustment or risk mitigation strategies.
2. **Vehicle Value vs. Total Claims, Colored by Loss Ratio and Sized by Premium (Scatter Plot):** *(Visual: Scatter plot of CustomValueEstimate vs TotalClaims, sized by TotalPremium, colored by LossRatio)* **Insight:** This multi-dimensional plot helps identify patterns where high-value vehicles might unexpectedly have low claims, or vice versa. It helps visualize if premiums are aligned with the custom value and actual claims, indicating potential opportunities for re-pricing specific segments of the market.
3. **Percentage Distribution of Insurance Cover Categories Over Time (Stacked Area Plot):** *(Visual: Stacked Area Plot of CoverCategory Distribution over Time)* **Insight:** This plot demonstrates shifts in customer preferences for different types of cover (`Comprehensive`, `Third Party`, `Windscreen`, etc.) over the 18-month period. For example, a decreasing trend in comprehensive cover might suggest economic pressures on clients, prompting ACIS to consider more affordable, specialized products or adjust marketing focus.

---

5. Recommendations for Plan Features and Marketing

Based on the preliminary EDA, the following recommendations are proposed:

5.1. Targeted Premium Adjustments & Marketing

- **Geographic Targeting (Provinces & Postal Codes):**
  - **Reduce Premiums:** For clients residing in `Province`s and `PostalCode`s identified with consistently **low Loss Ratios** . This segment can be marketed as "low-risk" targets, offering competitive pricing to attract new customers.
  - **Review/Increase Premiums:** For clients in areas with **high Loss Ratios** to ensure profitability, while simultaneously exploring risk mitigation strategies for these areas (e.g., promoting tracking devices, specific anti-theft measures).
- **Vehicle Type Specific Pricing:**
  - **Re-evaluate Premiums:** For `Heavy Commercial Vehicles` and other vehicle types with high average Loss Ratios.
  - **Competitive Offers:** For `Passenger Vehicles` or other types showing lower risk.
- **Vehicle Make/Model Focus:**
  - **Tailored Products:** Develop specific product features or discounts for makes/models associated with consistently low claim amounts.
  - **Risk Loading:** Implement risk loading for models with historically high claims.

5.2. Data-Driven Marketing Campaigns

- **Seasonal Campaigns:** Leverage insights from temporal trends (e.g., if claims spike in certain months, run campaigns promoting safer driving or specific covers before those periods).
- **Demographic-Specific Messaging:** While Gender differences were minor, if formal hypothesis testing confirms significance, tailor marketing messages. For instance, if one gender consistently shows lower risk, this could be a point of differentiation in marketing.
- **Product Feature Emphasis:** Based on the Cover Category trend, if certain covers are gaining or losing popularity, adjust marketing focus to either promote declining but profitable products or capitalize on growing segments.

5.3. Enhance Data Collection

- **Address Missing Gender Data:** Implement measures to reduce the "Not specified" category for `Gender` as this can provide valuable demographic insights. This could involve clarifying forms or communication during policy application.

- **Improve Citizenship/Marital Status Capture:** Investigate the reasons for high missingness in these fields, as they might offer additional segmentation opportunities.

5.4. Future Analytical Steps

- **A/B Hypothesis Testing:** Formally test the null hypotheses regarding risk differences across provinces, zip codes, and gender. This will provide statistical confidence for the observed differences.
- **Advanced Feature Engineering:** Create new features from existing data (e.g., `VehicleAge` from `RegistrationYear` or `VehicleIntroDate`, `ClaimFrequency` per policy).
- **Machine Learning Model Development:** Build predictive models for `TotalClaims` and `Optimal Premium` based on all available features, assessing the explaining power of important features.

---

6. Conclusion

The initial phase of this project has successfully laid the groundwork for advanced insurance analytics at ACIS. By deeply understanding the historical claim data, we have identified compelling patterns related to geography, vehicle characteristics, and temporal dynamics. These findings are directly actionable for refining ACIS's marketing strategies and identifying prime "low-risk" segments for competitive premium offerings. The next phases will build upon this foundation with rigorous statistical testing and predictive modeling to further enhance ACIS's capabilities in risk assessment and market optimization.