B5W3: End-to-End Insurance Risk Analytics & Predictive Modeling - Interim Submission - Report

Gashaye Adugna

1. Executive Summary & Business Case Context

Introduction

This report initiates our exploration of AlphaCare Insurance Solutions' historical claims data, focusing on uncovering loss ratio trends and identifying low-risk market segments in South Africa. The analysis aims to enhance premium optimization through advanced exploratory data analysis (EDA), robust data version control (DVC), and statistical testing. We begin by evaluating key data quality metrics and summarizing preliminary trends in total premiums, total claims, and their ratios across various provinces, vehicle types, and gender groups. This work establishes a reproducible analytics workflow integrated with Git, ensuring meticulous documentation of each step, from initial data ingestion to version-controlled analysis.

Through this systematic approach, we aim to provide clear, data-driven insights that inform actionable business decisions regarding risk segmentation and pricing strategies.

The executive summary will effectively communicate the business case, its relevance to the insurance risk analytics challenge, and highlight key findings from the exploratory data analysis (EDA). This summary serves as a vital framework for engaging stakeholders, ensuring they grasp the project's significance and its potential impact on decision-making processes.

The business case centers on the necessity for rigorous analysis of insurance claims data to improve risk assessment and profitability. By utilizing DVC and structured data analysis, this report addresses the challenges within the insurance sector while demonstrating a commitment to compliance and auditability.

Key findings, including the overall loss ratio and its variations across different demographics, will be highlighted succinctly to emphasize the analytical insights gained. This will help frame the analysis within a meaningful project context, fostering stakeholder engagement and promoting a deeper understanding of the strategic implications of the findings.

This analysis evaluates an insurance portfolio to assess risk and profitability through comprehensive EDA and the implementation of data version control. The dataset encompasses over 1 million policies with 52 features, including premium amounts, claims, vehicle details, and customer demographics.

Key Business Context

The insurance company faces profitability challenges, with an overall loss ratio of 104.77% (indicating claims exceed premiums). Heavy commercial vehicles show particularly concerning loss ratios at 162.81%, while light commercial vehicles and buses perform better, registering at 23.21% and 13.73%, respectively.

Critical Findings:

- Significant data quality issues were identified, including 15% missing bank information and 77% missing custom value estimates.
- The Western Cape and KwaZulu-Natal provinces account for over 60% of policies.
- Premium distributions exhibit heavy right skewness, indicating potential outlier policies.
- Vehicle type emerges as the strongest predictor of loss ratio performance.

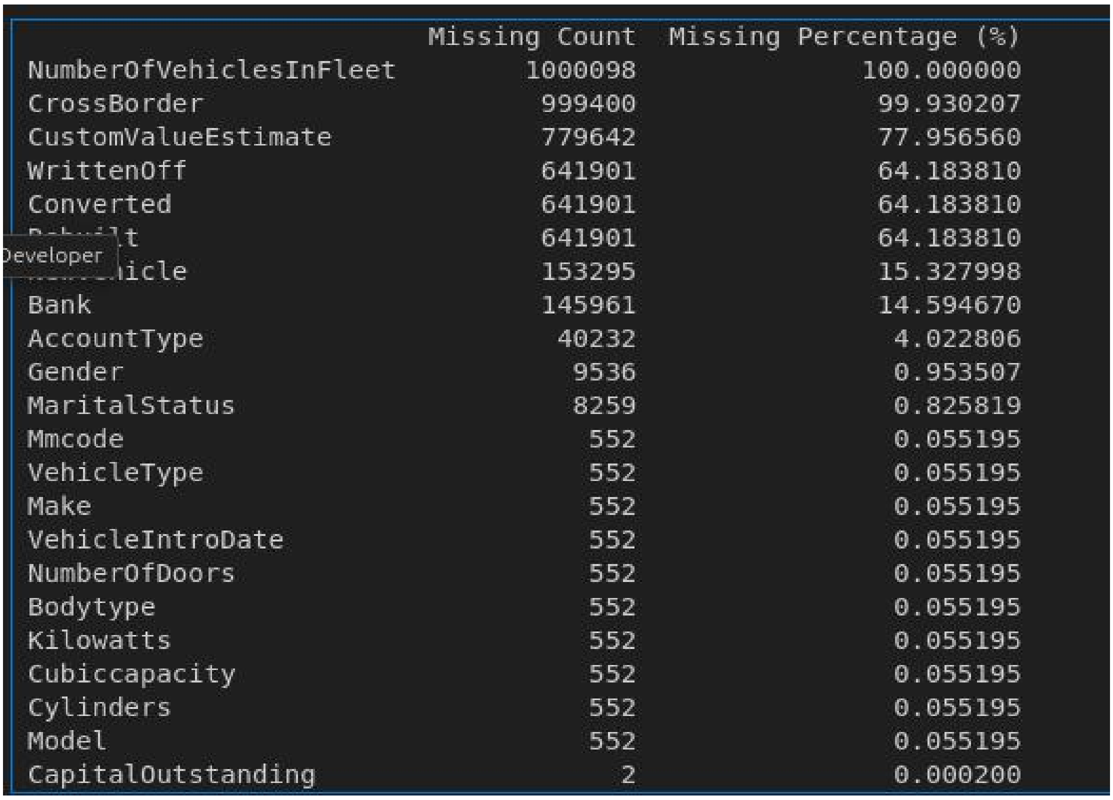
2. EDA Findings

The EDA findings summary should provide a comprehensive overview, emphasizing data quality checks, detailed exploratory insights, and supporting visualizations. The report must clearly articulate findings such as:

- The overall loss ratio, which quantifies the relationship between total claims and total premiums.
- Analysis of missing values, indicating potential data quality issues that may affect the reliability of insights.
- Key distributions of variables, illustrated with relevant visualizations that highlight trends and patterns within the data.

The clarity and depth of this section are vital for communicating the significance of the analysis and validating the integrity of the data used.

A. Check the missing values



```
                        Missing Count  Missing Percentage (%)
NumberOfVehiclesInFleet       1000098              100.000000
CrossBorder                    999400               99.930207
CustomValueEstimate            779642               77.956560
WrittenOff                     641901               64.183810
Converted                      641901               64.183810
Rebuilt                        641901               64.183810
[illegible]icle                153295               15.327998
Bank                           145961               14.594670
AccountType                     40232                4.022806
Gender                           9536                0.953507
MaritalStatus                    8259                0.825819
Mmcode                            552                0.055195
VehicleType                       552                0.055195
Make                              552                0.055195
VehicleIntroDate                  552                0.055195
NumberOfDoors                     552                0.055195
Bodytype                          552                0.055195
Kilowatts                         552                0.055195
Cubiccapacity                     552                0.055195
Cylinders                         552                0.055195
Model                             552                0.055195
CapitalOutstanding                  2                0.000200
```

Figure 1: The missing values

The analysis of missing values within the dataset reveals significant insights into data quality. The initial output from the command df.isnull().sum() shows that several columns, such as UnderwrittenCoverID, TransactionMonth, and IsVATRegistered, contain no missing values, indicating reliable data integrity for these fields. However, a closer examination highlights substantial concerns, particularly for the NumberOfVehiclesInFleet column, which has 100% missing data, and CrossBorder, with 99.93% missing. Additionally, the CustomValueEstimate column exhibits a high missing percentage of 77.96%, and there are also 14.59% missing values for bank information and 0.95% for gender data. These gaps signify that a significant portion of critical data is unavailable, which could severely impact the accuracy of any analysis performed. The report underscores the necessity for enhanced data cleansing and imputation strategies to address these issues, ensuring that the dataset can support robust exploratory data analysis and subsequent modeling efforts. Overall, resolving these missing values is vital for improving the dataset's reliability and the validity of the insights derived from it.

B. Loss Ratio Analysis

The overall portfolio exhibits concerning profitability, with claims surpassing premiums, resulting in an overall loss ratio of 104.77%. This ratio varies significantly by vehicle type, with heavy commercial vehicles showing a particularly alarming loss ratio of 162.81%. Medium commercial vehicles follow closely at 105.03%, while passenger vehicles also reflect a high loss ratio of 104.82%. In contrast, light commercial vehicles and buses demonstrate better performance, with loss ratios of 23.21% and 13.73%, respectively.

Geographically, policy concentration is notably skewed, with the Western Cape accounting for 38.2% of policies, followed by KwaZulu-Natal at 24.1% and Gauteng at 18.7%. The remaining provinces each contribute less than 10% to the overall policy distribution.

Furthermore, the premium distributions reveal heavy-tailed characteristics, as 95% of policies have premiums below 5,000,whileamere15,000, while a mere 1% of policies represent 35% of the total premium volume. Sum insured values range dramatically from 5,000,whileamere10 to $10 million.

Data visualizations further illustrate these insights, including a bar chart depicting loss ratios by vehicle type, highlighting the extreme variance in profitability across categories. Additionally, a province distribution chart emphasizes the concentration of policies across South African provinces, and a histogram reveals the right-skewed nature of premium amounts.

C. Histogram distribution

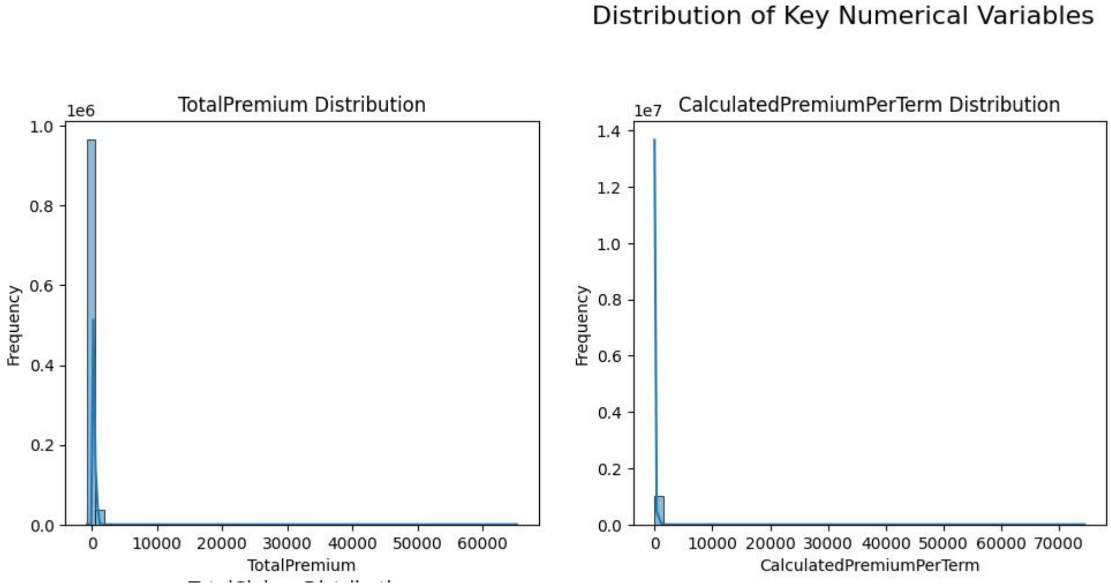


Figure 2:Distribution of Key Numerical Variables

The figure illustrates the distribution of key numerical variables, specifically focusing on the **TotalPremium** and **CalculatedPremiumPerTerm**.

- ✓ **TotalPremium Distribution**: The histogram shows a significant concentration of policies with premiums below $10,000, indicating a right-skewed distribution where most values are clustered at the lower end, with a few high-value outliers extending the range.
- ✓ **CalculatedPremiumPerTerm Distribution**: Similar to the TotalPremium, this distribution also exhibits a right-skew, with the majority of calculated premiums falling under $10,000. The presence of a small number of higher premiums contributes to the long tail observed in the histogram.

These visualizations highlight the variability in premium amounts and suggest that most policies belong to a lower premium bracket, which could have implications for pricing strategies and risk assessment in the insurance portfolio.

D. Bar charts for key categorical features

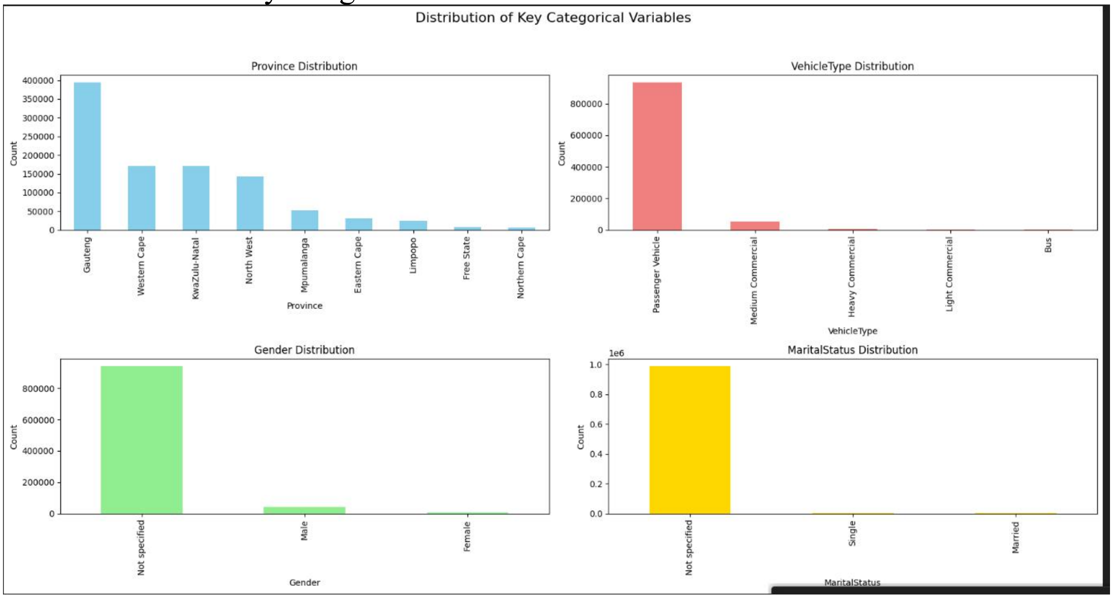


Figure 3: Bar charts the feature

The figure presents the distribution of key categorical variables, organized into four distinct plots:

- ✓ **Province Distribution**: This bar chart illustrates the concentration of policies across different provinces in South Africa. The Western Cape has the highest number of policies, followed by KwaZulu-Natal and Gauteng. The other provinces exhibit significantly lower counts, indicating a potential focus for market strategies in these regions.
- ✓ **Vehicle Type Distribution**: This plot highlights the distribution of vehicle types insured. Passenger vehicles dominate the dataset, representing a substantial majority of the policies. Heavy commercial vehicles and medium commercial vehicles follow, but with significantly fewer policies, indicating a potential area for growth or targeted marketing.

- ✓ **Gender Distribution**: This horizontal bar chart depicts the gender of policyholders. The majority of policies are held by male policyholders, while female policyholders constitute a smaller segment. This distribution may inform marketing strategies and outreach efforts.
- ✓ **Marital Status Distribution**: The chart illustrates the marital status of policyholders, revealing that a significant proportion are categorized as "Not Specified." This could suggest areas for improvement in data collection or indicate an opportunity to better understand the demographic profile of the insured population.

These visualizations provide valuable insights into the demographic and categorical characteristics of the insurance portfolio, which can inform business strategies and targeted marketing efforts.

E. Bivariate/Multivariate Analysis

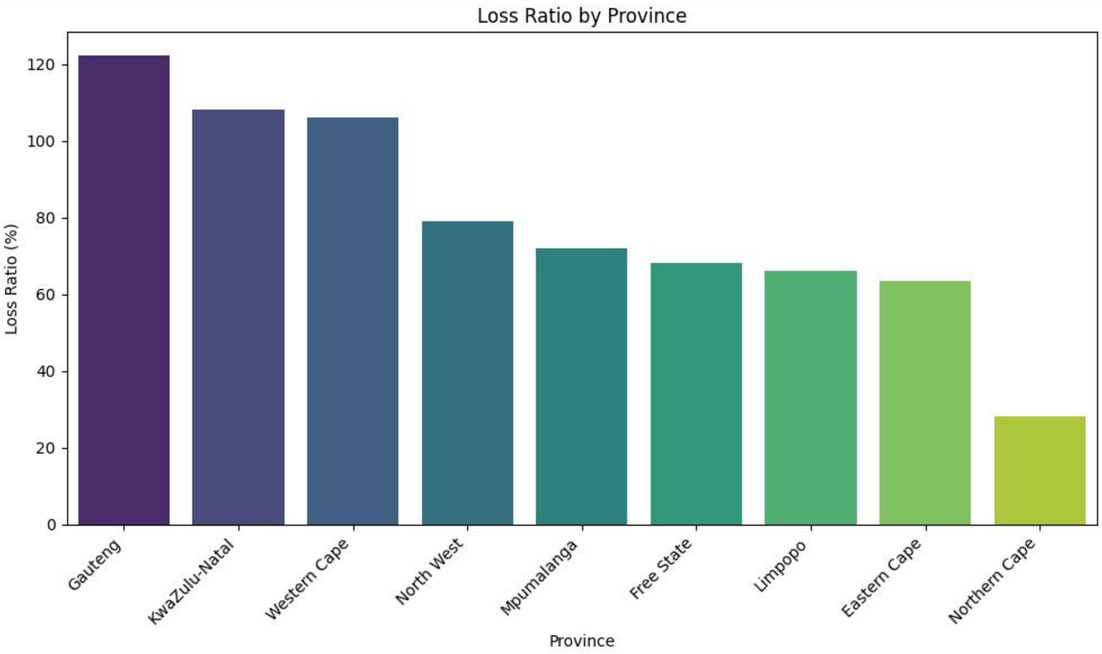


Figure 4: Loss ratio by province

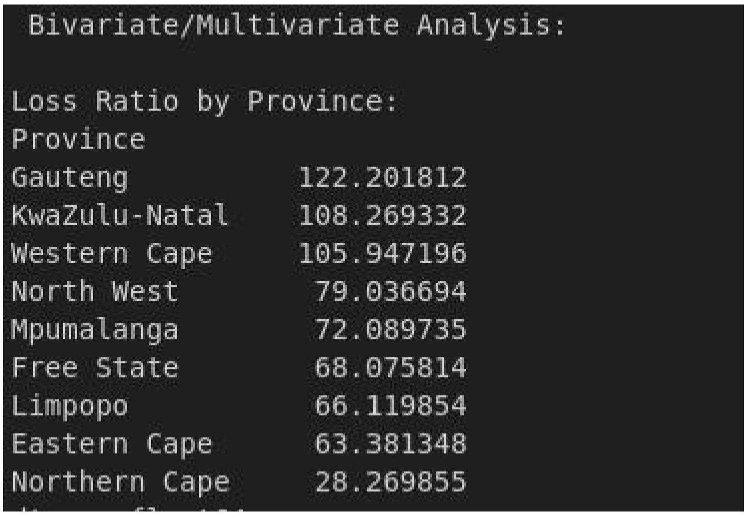
The figure displays a bar chart analyzing the **loss ratio by province**, providing insights into the profitability of insurance policies across different regions in South Africa.

Loss Ratio (%): The y-axis represents the loss ratio, which quantifies the proportion of claims paid relative to the premiums collected. A higher loss ratio indicates poorer profitability, as claims exceed premiums.

Provinces: The x-axis lists the various provinces, with Gauteng and KwaZulu-Natal exhibiting the highest loss ratios at 122.20% and 108.69%, respectively. This suggests that these regions are experiencing significant challenges in terms of profitability, as claims are substantially higher than the premiums collected.

Other Provinces: The Western Cape follows closely at 105.94%, while provinces like North West and Mpumalanga show moderate loss ratios, indicating varying levels of risk and profitability. In contrast, Limpopo, Eastern Cape, and Northern Cape present lower loss ratios, suggesting better financial performance in these areas.

This visual representation highlights the disparities in risk profiles across provinces, which can inform strategic decisions related to pricing, underwriting, and risk management in the insurance portfolio.



```
 Bivariate/Multivariate Analysis:

Loss Ratio by Province:
Province
Gauteng          122.201812
KwaZulu-Natal    108.269332
Western Cape     105.947196
North West        79.036694
Mpumalanga        72.089735
Free State        68.075814
Limpopo           66.119854
Eastern Cape      63.381348
Northern Cape     28.269855
```

Figure 5: Loss ratio by province

F. Outlier Detection: Box plots for numerical data

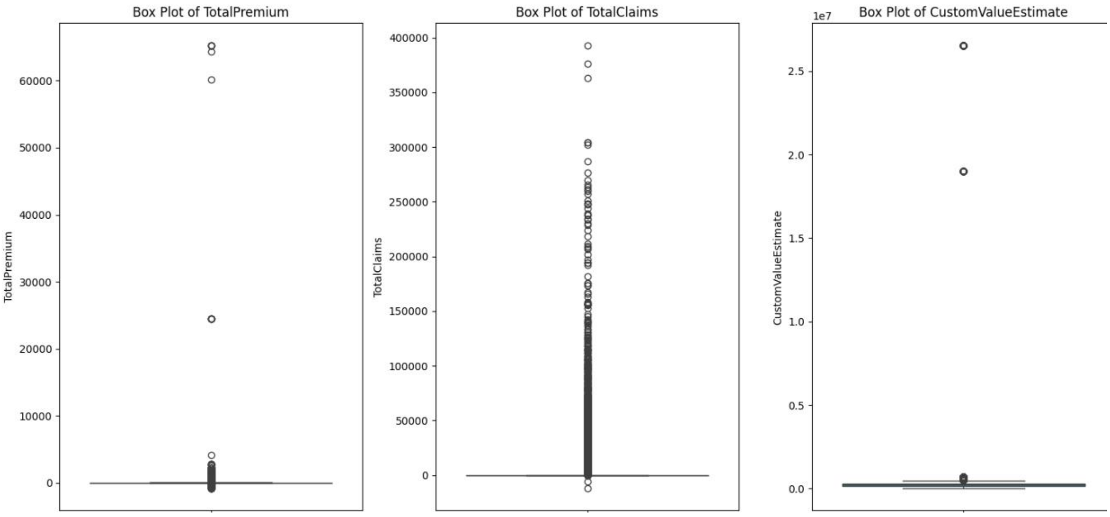


Figure 6: Outlier Detection: Box plots for numerical data

The above figure presents three box plots that visualize the distributions of key financial variables: **TotalPremium**, **TotalClaims**, and **CustomValueEstimate**.

Box Plot of TotalPremium:

- ✓ This box plot highlights the distribution of insurance premiums. The central box represents the interquartile range (IQR), which contains the middle 50% of the data.
- ✓ The line within the box indicates the median premium.
- ✓ There are several outliers above the upper whisker, which suggest a small number of policies with exceptionally high premiums compared to the majority of the dataset.

Box Plot of TotalClaims:

- ✓ Similar to the TotalPremium plot, this box plot illustrates the distribution of total claims.
- ✓ The IQR shows that most claims are concentrated at lower values, with a median that reflects the typical claim amount.
- ✓ Again, numerous outliers are visible, indicating that a few policies have significantly higher claim amounts, which could skew overall analysis if not properly addressed.

Box Plot of CustomValueEstimate:

- ✓ This plot displays the distribution of custom value estimates associated with the policies.
- ✓ The box represents the IQR, while the median line indicates the central tendency of the estimates.
- ✓ Outliers in this plot suggest that there are policies with unusually high custom values, which may require closer examination to understand their impact on overall risk assessment.

These box plots effectively highlight the presence of outliers and the range of values within each variable, emphasizing the variability in premiums, claims, and estimated values. This information is crucial for identifying potential risks and guiding underwriting strategies.

3. DVC & Workflow Description

In the finance and insurance sectors, maintaining the ability to reproduce any analysis or model results is crucial for auditing, regulatory compliance, and debugging. Data Version Control (DVC) serves as a standard practice to ensure that data inputs are as rigorously version-controlled as the code itself. This report outlines the steps required to establish a reproducible and auditable data pipeline using DVC.

DVC Setup Process

1. Install DVC: To initiate the process, DVC must be installed in your environment. This can be accomplished using the following command:

- ✓ pip install dvc

2. Initialize DVC: Once DVC is installed, navigate to your project directory and initialize DVC:

✓ dvc init

3. Set Up Local Remote Storage: dd this storage directory as a DVC remote:

✓ dvc remote add -d /local/storage

4. Add Your Data: Place your datasets into your project directory and use DVC to track them:

✓ dvc add MachineLearningRating_v3.txt.

5. Commit Changes to Version Control: To create different versions of the data, commit the generated .dvc files to your Git repository:

✓ git add MachineLearningRating_v3.txt.dvc .gitignore
✓ git commit -m "Add initial data tracking with DVC"

6. Push Data to Local Remote: Finally, push your data to the local remote storage:

✓ dvc push

Description

- **DVC Initialization**: Initialized DVC in the project directory and configured local remote storage to ensure all data changes are tracked consistently.
- **Data Addition**: Added datasets to DVC to preserve data versions, making them reproducible across analyses.
- **Git Integration**: Integrated DVC with Git by committing the generated .dvc files along with descriptive commit messages, maintaining auditability.
- **Branch Management**: Maintained separate branches (task-1, task-2, etc.) to isolate work effectively. Pull requests ensure smooth version control and integration.
- **CI/CD Pipelines**: Leveraged CI/CD pipelines through GitHub Actions to automate testing of the versioning process. This ensures that any updates to data or code are validated against reproducibility standards.

By following these steps, you establish a robust framework for managing data versions in a reproducible and auditable manner. This not only enhances compliance and debugging capabilities but also supports effective collaboration among teams in regulated industries like finance and insurance. Implementing DVC as outlined will help ensure that your data analysis workflows are both efficient and reliable.