Data Mining in Actuarial Science: Reserves

Following the CRISP-DM methodology.

Juan Lara

November 9, 2023

Disclaimer

The information contained in this document is provided for general informational purposes only and does not constitute professional advice, financial guidance, or any other form of recommendation. While we have taken every effort to ensure the accuracy and reliability of the information presented, Universidad Nacional de Colombia makes no representations or warranties of any kind, express or implied, about the completeness, accuracy, reliability, suitability, or availability of the information, products, services, or related graphics contained in this document for any purpose.

Any reliance you place on such information is strictly at your own risk. In no event will Universidad Nacional de Colombia be liable for any loss or damage including without limitation, indirect or consequential loss or damage, or any loss or damage whatsoever arising from loss of data or profits arising out of, or in connection with, the use of this document.

Through this document, you may be able to link to other websites that are not under the control of Universidad Nacional de Colombia. We have no control over the nature, content, and availability of those sites. The inclusion of any links does not necessarily imply a recommendation or endorse the views expressed within them.

Every effort is made to keep the document up and running smoothly. However, Universidad Nacional de Colombia takes no responsibility for, and will not be liable for, the document being temporarily unavailable due to technical issues beyond our control.

Before making any business decisions or taking any actions based upon the information provided in this document, we recommend consulting with appropriate professionals or advisors. The content of this document is subject to change without notice.

By using this document, you agree to the terms outlined in this disclaimer.

Copyright

Contact

jlara@unal.edu.co

Changelog

v0.1	2023-08	Business Understanding: Introduced reserves and objectives, assessed risks, proposed a central machine learning model, detailed "ActuaSolutions" structure.
v0.3	2023-09	Data Understanding:
v0.5	2023-09	Data Preparation:
v0.9	2023-10	Modeling:
v0.9.1	2023-10	Evaluation:
v1	2023-10	Deployment:

Table of Contents

1 Business Understanding

Actuarial reserves, especially in the context of insurance, are estimates of future obligations that a company has due to already issued insurance policies. These estimates consider uncertain events, such as future claims, and aim to ensure that the company has the necessary funds to fulfill its responsibilities as they arise. Following the instructions in [2] we process describing the objectives.

1.1 Business Objectives

Proper management and calculation of reserves are essential to maintain the financial health of **ActuaSolutions** [1] and ensure its strength in the market. By refining this process, the company aims to:

- Ensure solvency and the ability to meet all future obligations.
- Strengthen the trust of its clients, shareholders, and regulators.
- Optimize the allocation of financial resources.
- Establish itself as a leader in the actuarial field by adopting innovative and accurate methodologies.

1.1.1 Organizational Structure

The company "ActuaSolutions" is a leading organization in the actuarial field that has firmly established itself in the market due to its commitment to accuracy, innovation, and adaptability. To better understand how the company operates, it is essential to grasp its organizational structure and how different departments collaborate to achieve business objectives.

At the heart of our actuarial company is a deep understanding of our structured organization which influences our approach to challenges and solutions:

- *Organizational Charts:* Our corporate landscape is detailed with divisions, departments, and project groups. This chart encompasses the hierarchy from top management to the operational level, detailing managers, their respective roles, and responsibilities.
- *Key Individuals:* We've recognized pivotal personnel across departments whose expertise and insights drive our operations forward. Their roles range from strategic planning, operations, to domain-specific expertise in actuarial science.
- *Internal Sponsorship:* Our internal sponsor bridges the financial and domain expertise, ensuring that the project receives both the necessary funds and knowledge backup.
- *Steering Committee:* Our committee provides direction and oversight. Comprising a diverse group of experts, this team ensures that the project aligns

with our corporate goals.

- *Affected Business Units:* Given the cross-functional nature of data mining, multiple business units including marketing, finance, and operations would be stakeholders in the project's outcome.

Interdepartmental collaboration is essential for the company to maintain its excellence and remain a leader in the actuarial field. We describe each department as:

1. **Actuarial Reserves Department:**

   Functionality: This is the heart of the company. This department is responsible for calculating reserves, estimating future obligations, and ensuring that the company has the necessary resources to fulfill its responsibilities.

   Key Teams:

   - *Actuarial Analysts:* Specialized professionals who use mathematical and statistical techniques to assess risk and determine reserves.
   - *Data Specialists:* Experts who work with large datasets, perform exploratory analyses, and prepare data for analysis.
   - *IT Team:* Responsible for maintaining and optimizing the tools and platforms used for actuarial calculations.

2. **Financial Department:**

   Functionality: Manages the company's financial resources, ensures obligations are met, and works closely with the actuarial department to adjust reserves.

   Key Teams:

   - *Accounting:* Monitors and reports financial flows.
   - *Investment Management:* Optimizes the company's investment returns.

3. **Claims Department:**

   Functionality: Manages and processes claims submitted by policyholders.

   Key Teams:

   - *Claims Adjusters:* Analyze and assess the validity of submitted claims.
   - *Customer Support Team:* Assist policyholders and facilitate the claims process.

4. **Development and Strategy Department:**

   Functionality: Responsible for researching new markets, products, and strategies to expand and strengthen the company's position.

   Key Teams:

   - *Market Research:* Study market trends and identify opportunities.
   - *Product Development:* Create and optimize the insurance products offered by the company.

5. **Technology and Innovation Department:**

   Functionality: Explores and adopts emerging technologies, develops software solutions, and seeks ways to implement artificial intelligence and machine learning in the company's processes.

   Key Teams:

   - *Developers:* Build and maintain technological platforms and tools.
   - *Machine Learning Specialists:* Design and train machine learning models for various purposes, including reserves.

Each department, with its unique teams and expertise, plays a vital role in the functioning of "ActuaSolutions".

1.1.2 Problem Area Identification

Our prime focus rests on the provision and actuarial domain, with the motivation to ensure accurate reserve estimations and enhanced financial health:

- **Domain:** Our challenge primarily lies within the actuarial realm, impacting areas such as risk assessment, financial obligations, and provisions.
- **Problem Description:** Our business faces the intricate challenge of accurately estimating future obligations based on past and current data.
- **Project Prerequisites:** This venture was conceived from the need for more precise provision estimations. Currently, while we utilize data-driven strategies, there is a gap in implementing advanced data mining techniques.
- **Status & Advocacy:** The project is in its nascent stage and requires both validation and advocacy. Recognizing this, we're ready to present the merits of data mining as a transformative solution to our board and stakeholders.

1.1.3 Current Solution Analysis

While we have legacy systems and processes in place, understanding their strengths and limitations is key:

- **Existing Solutions:** Our current strategy employs traditional actuarial methods combined with basic statistical tools to estimate provisions.
- **Advantages & Disadvantages:** While these methods have served us well over the years, providing consistency and a moderate degree of accuracy, they lack the adaptability and precision offered by advanced data mining. Furthermore, there's been a mixed reception within the organization about the current methods' effectiveness.

Understanding the above components provides a robust framework for our data mining project, ensuring that we're not just data-informed but also business-aligned.

1.2 Project Success Criteria

For a project as comprehensive and essential as ours, defining success is paramount. Recognizing this, we've broken down the criteria into two segments to capture the tangible and the intangible markers of accomplishment:

- **Accuracy Enhancement:** Achieve a specific percentage increase in the accuracy of our reserve estimations compared to our current methods.
- **Operational Efficiency:** Reduce the time taken for provision calculations by an agreed-upon percentage.
- **Financial Improvement:** Realize a quantifiable reduction in financial discrepancies or shortfalls due to more accurate provision calculations.

1.2.1 Subjective Success Criteria

These criteria, while not strictly quantifiable, are essential to ensure that our solutions align with the broader organizational vision and stakeholder expectations:

- **Enhanced Decision-making:** Uncover patterns and insights from our data that empower decision-makers to strategize more effectively.
- **Stakeholder Satisfaction:** Achieve a consensus among key stakeholders that the data mining model provides valuable and actionable insights. Here, feedback from department heads and the steering committee would be pivotal.
- **Discovery of Effective Actuarial Techniques:** While this is harder to pin down, our aim is to uncover and integrate innovative actuarial methods or clusters of techniques that resonate with our company's challenges and goals.

1.3 Assesing the Situation

To set the project on a solid foundation, it's crucial to understand our current standing, the resources we possess, and potential challenges we might face. The following assessment provides a comprehensive understanding of our present resources, needs, constraints, risks, and potential rewards.

1.3.1 Resource Inventory

Before plunging into the data mining process, it's imperative to evaluate our resources thoroughly:

- **Hardware Resources:** An assessment of the necessary hardware to support the data mining processes is vital. This encompasses server capabilities, storage, and other infrastructural necessities.
- **Data Sources and Knowledge Stores:**

  - We have various data sources, each with its own type and format. It's essential to note these details as they dictate the preprocessing steps required.
  - Currently, our data is stored in a combination of data warehouses and operational databases. We do have live access to these, ensuring real-time data analysis.
  - External data acquisition, like demographic details, is under consideration. The costs and benefits will determine this decision.
  - Security concerns, especially with sensitive data, have been flagged. We're ensuring that data access adheres to the company's privacy and security guidelines.
- **Personnel Resources:** Our team comprises business and data experts, database administrators, and additional support staff. They're the backbone of the project, ensuring seamless execution and valuable insights.

1.3.2 Requirements, Assumptions, and Constraints

Transparent understanding and communication regarding the project's boundaries and needs are pivotal:

- **Requirements:**
  - Adhering to legal and security restrictions on the data and results is a top priority.
  - Project scheduling has been communicated and agreed upon by all stakeholders.
  - Result deployment methods, including potential web publishing or database entries, have been outlined.
- **Assumptions:**
  - The project may incur additional costs, such as consulting fees, which have been factored into our projections.
  - Data quality has been assumed to a certain standard. Any deviation might necessitate additional preprocessing steps.
  - The management team primarily expects result-oriented insights, but a deeper dive into the models can be provided upon request.
- **Constraints:**
  - All necessary data access permissions have been obtained.
  - Legal constraints on data usage have been vetted.
  - The project budget encompasses all financial requirements.

1.3.3 Risks and Contingencies

Every venture has its potential pitfalls, and a key risk we must address is the possibility of not acquiring the necessary data for model formulation. Identifying and

mitigating these risks in advance is pivotal to ensuring a seamless and unhindered progression:

- **Risks:**
  - Scheduling: Project delays.
  - Financial: Budgetary concerns from the project sponsor.
  - Data: Quality or coverage issues.
  - Results: Not meeting initial expectations.
  - Data Availability: Challenges in obtaining requisite data for model development.
- **Contingency Plans:**
  - For **Scheduling Risk:** Regularly review project milestones, identify potential delays in advance, and adjust the project plan accordingly.
  - For **Financial Risk:** Maintain open communication with the project sponsor, establish a buffer in the budget for unforeseen expenses, and seek alternative funding sources if needed.
  - For **Data Risk:** Implement data quality assessment measures, invest in data cleansing and enrichment, and explore alternative data sources if necessary.
  - For **Results Risk:** Continuously validate and recalibrate the model, involve stakeholders in setting realistic expectations, and prioritize transparency in reporting results.
  - For **Data Availability Risk:** Establish multiple data procurement channels, collaborate closely with data providers, and consider using synthetic data as a temporary substitute.

1.3.4 Cost/Benefit Analysis

An in-depth analysis to ascertain the potential return on our investment:

- **Costs:**
  - Data collection and any external data.
  - Result deployment.
  - Operational expenses.
- **Benefits:**
  - Achieving the primary objective.
  - Gaining deeper insights from data exploration.
  - Advantages arising from a superior understanding of the data.

This comprehensive assessment not only lays the groundwork for our data mining project but also ensures that we are always cognizant of our strengths, challenges, and the value we aim to create.

1.4 Data Mining Objectives

The primary purpose is to develop a machine-learning model that can predict future obligations more accurately based on historical data and other relevant variables. This model seeks to identify complex patterns that traditional methodologies might overlook. Furthermore, the proposed model is expected to be adaptable to changes in the market environment and can be retrained and adjusted as needed.

1.5 Data Mining Project Plan

Given the pivotal role that a project plan plays in aligning team members, stakeholders, and resources to the project's goals, the plan must be comprehensive, accurate, and well-informed. Drawing inspiration from the guidelines shared, here is a meticulously crafted project plan for our data mining endeavor.

1.5.1 Project Plan Description

The project plan, which will serve as our central document, encapsulates our goals, resources, risks, and timelines for every phase of the data mining process. This ensures that everyone, from stakeholders to team members, is well-informed about the project's progress and objectives. To enhance accessibility and increase collaboration, we propose publishing this plan on the company's intranet.

The project plan creation involved discussions with all the stakeholders to ensure its feasibility. We have made sure to include time estimates for all tasks, highlight decision points, and emphasize stages that might require multiple iterations.

Phase	Time (Days)	Resources	Risks
Business Understanding	7	All analysts	Economic changes
Data Understanding	21	All analysts	Data-related issues, technological problems
Data Preparation	35	Data mining consultant, some database analyst time	Data-related issues, technological problems
Modeling	14	Data mining consultant, some database analyst time	Technological issues, finding a suitable model
Evaluation	7	All analysts	Economic changes, challenges in result implementation
Deployment	7	Data mining consultant, some database analyst time	Economic changes, challenges in result implementation

Table 1: Data Mining Project Plan

2 Data Understanding

2.1 Collecting Initial Data

Within our CRISP-DM methodology, our data journey begins with curating a pivotal dataset sourced from the Loss Reserving Data of the National Association of Insurance Commissioners (NAIC) Schedule P. This dataset, meticulously compiled for claims reserving studies, originates from a wealth of claims data encompassing major personal and commercial lines in the US property casualty insurance sector. Comprising six distinct lines of business, from private passenger auto liability to workers' compensation, this repository unveils the intricate tapestry of risk management and coverage within the diverse insurance landscape.

The data preparation process involved a meticulous three-step approach, ensuring that the dataset was refined and robust for our analytical endeavors:

Step I: Pulling Triangle Data: Our journey begins with the extraction of triangle data from the Schedule P of the year 1997. Each of these triangles encapsulates claims spanning 10 accident years from 1988 to 1997, coupled with 10 development lags. This initial dataset forms the bedrock of our model development phase.

Step II: Triangle Squaring: In the pursuit of enhancing our model's efficacy, we square the triangles from the 1997 Schedule P dataset with outcomes from subsequent years. This cross-temporal integration allows us to validate and test our model's performance retrospectively.

Step III: Quality Assurance through Sampling: The dataset's integrity is of paramount importance. To ensure its quality, a meticulous sampling process was conducted. Insurers were retained in the final dataset based on stringent criteria. These include the completeness of observations, alignment of claims data, and non-zero net premiums written for all years.

The culmination of this thorough data preparation process is a refined dataset, encapsulating run-off triangles spanning six distinct lines of business. These triangles encapsulate claims from the accident year 1988 to 1997, paired with a 10-year development lag. Importantly, both upper and lower triangles are included, affording us the opportunity to model and test the performance of our strategies with a retrospective lens.

2.2 Data Description

The dataset for our claims reserving studies project is derived from the Loss Reserving Data provided by the National Association of Insurance Commissioners (NAIC) through their Schedule P. This meticulously curated dataset encompasses a total of 13,200 records, each containing 13 attributes, making it a rich and comprehensive source of insights.

The dataset's diverse attributes and comprehensive size make it an invaluable resource for our claims reserving studies, enabling us to analyze historical patterns and trends within the Workers' compensation line of business.

- **GRCODE:** NAIC company code, including insurer groups and single insurers.
- **GRNAME:** NAIC company name, including insurer groups and single insurers.
- **AccidentYear:** Accident year (1988 to 1997).
- **DevelopmentYear:** Development year (1988 to 1997).
- **DevelopmentLag:** Development year (AY-1987 + DY-1987 - 1).
- **IncurLoss_D:** Incurred losses and allocated expenses reported at year end.
- **CumPaidLoss_D:** Cumulative paid losses and allocated expenses at year end.
- **BulkLoss_D:** Bulk and IBNR reserves on net losses and defense and cost containment expenses reported at year end.
- **EarnedPremDIR_D:** Premiums earned at incurral year - direct and assumed.
- **EarnedPremCeded_D:** Premiums earned at incurral year - ceded.
- **EarnedPremNet_D:** Premiums earned at incurral year - net.
- **Single:** 1 indicates a single entity, 0 indicates a group insurer.
- **PostedReserve97_D:** Posted reserves in year 1997 taken from the Underwriting and Investment Exhibit – Part 2A, including net losses unpaid and unpaid loss adjustment expenses.

Attributes such as 'GRCODE' and 'GRNAME' provide information about NAIC company codes and names. 'AccidentYear' and 'DevelopmentYear' span the years 1988 to 1997. Additional attributes like 'IncurLoss_D', 'CumPaidLoss_D', and others denote specific loss and premium data for the 'Workers' compensation' line of business.

2.3 Exploring Data

In this Data Exploration section, we present a comprehensive visualization of the key statistical characteristics of the run-off triangles, grouped by the indicators BulkLoss_D, CumPaidLoss_D, and IncurLoss_D. Each row in Figure 1 reflects one of these indicators, breaking down the data into measures of central tendency and variability. The first column displays the mean of the consolidated datasets, offering an overview of the average value across all triangles in our database. The second column exhibits the variance, providing quantification of data variability or dispersion. Finally, the third column presents the median, facilitating an understanding of the data distribution by indicating the central value. This systematic arrangement allows for direct and efficient comparison between different statistical metrics, essential for accurate interpretation of the underlying dynamics in actuarial reserve data.

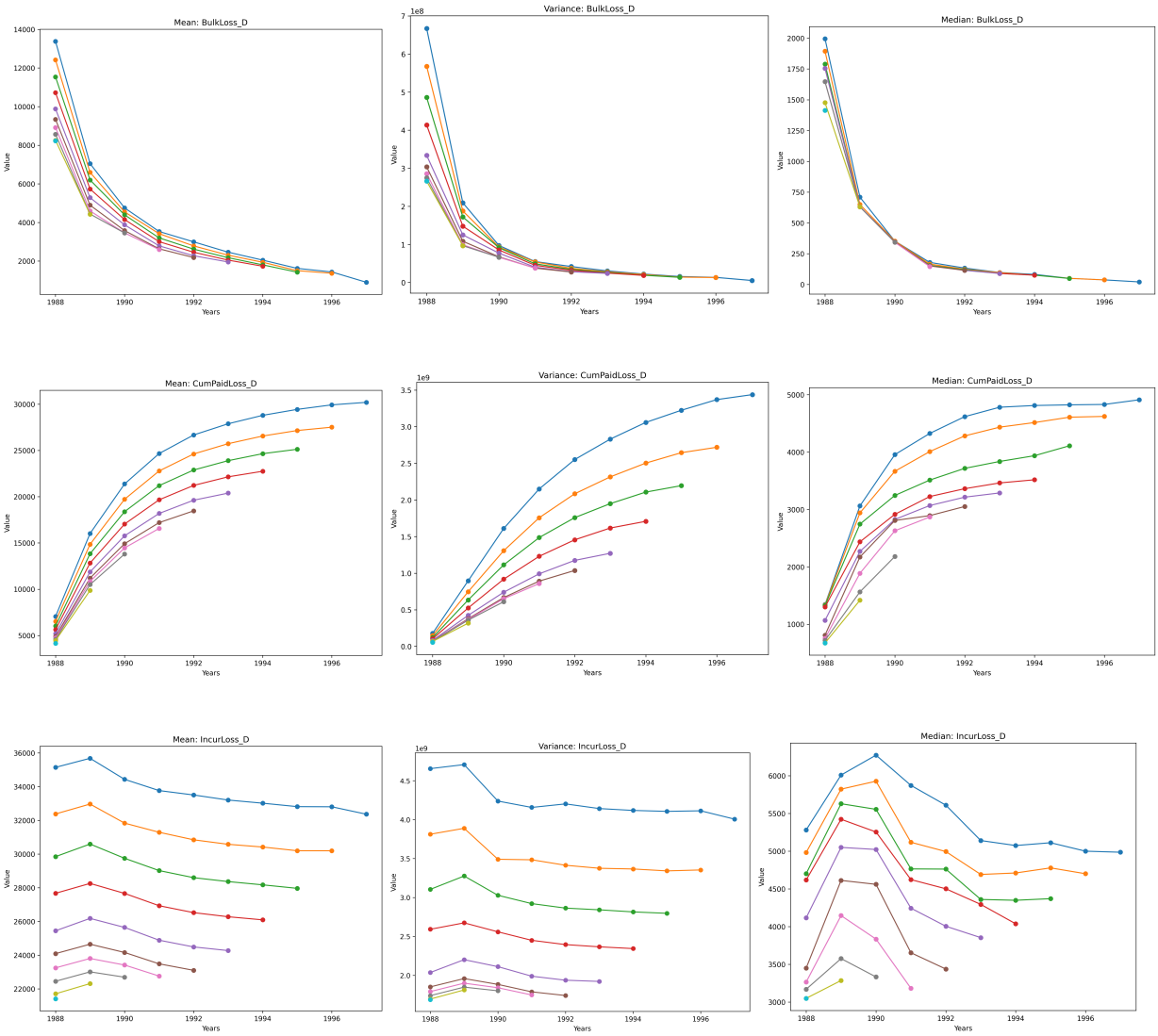

Figure 1: Statistical Overview of Run-Off Triangle Metrics

The preceding Figure 1 visually synthesizes the essential statistical information of the run-off triangles, providing a solid foundation for further analysis in the process of Actuarial Data Mining. The orderly layout in a 3x3 format emphasizes the differences and similarities between the measures of mean, variance, and median, which are crucial for actuarial valuation and reserve prediction. The detailed graphical exploration of BulkLoss_D, CumPaidLoss_D, and IncurLoss_D paves the way for an in-depth discussion on the suitability of predictive models and the robustness of reserve estimates. The insights gained from these visualizations will be of paramount importance in the subsequent phases of the “Data Mining in Actuarial Science: Reserves" project, particularly when addressing the selection and application of advanced analytical techniques within the CRISP-DM methodology. The interpretation of these figures will contribute to building a comprehensive framework for strategic decision-making in actuarial risk management.

3 Data Preparation

The Data Preparation phase in our claims reserving actuarial project is instrumental in transforming raw data into an analytical framework. This phase is segmented into systematic tasks, each designed to refine the dataset, culminating in a form conducive to advanced actuarial analysis.

3.1 Variable Selection

Identifying critical variables constitutes the initial step in our data preparation process. We meticulously extracted key attributes—namely incurred losses, paid losses, bulk reserves, and earned premiums—from the dataset. These attributes are pivotal in examining the financial dynamics of claims over time and were chosen for their relevance to our focused study.

3.2 Data Structuring

Subsequent to variable selection, we embarked on structuring the data into matrices, representing a decade of claims development for each variable. This 10x10 matrix format is crucial for visualizing the progression of claim settlements and reserve allocations throughout the development period.

3.3 Triangulation

Triangulation is a cornerstone of actuarial data preparation, essential for the application of reserving methods such as the chain-ladder and Bornhuetter-Ferguson techniques. We transformed our structured matrices into triangular forms, in alignment with actuarial conventions, to facilitate the modeling of claims payment development and reserve estimation.

3.4 Initial Data Check

Upon loading the dataset, we verified the absence of missing values and ensured the proper assignment of data types to each variable. This verification step is a precursor to further data refinement activities.

3.5 Addressing Negative Values

We identified and assessed the presence of negative values within the financial columns, attributing them to potential adjustments or returns in financial records. These values were preserved in the dataset to maintain the integrity of our actuarial analysis.

3.6 Normalization

Prior to dimensionality reduction techniques, we performed normalization of numerical columns. This step is crucial for equalizing the contribution of each variable to the analysis, thereby preventing any skewness due to variance in variable scales.

3.7 Dimensionality Reduction with PCA

Applying Principal Component Analysis, we achieved a reduction in dataset dimensionality, identifying principal components that encapsulate a significant proportion of the dataset's variance. The PCA revealed that the first two components account for approximately 89% of the total variance, indicating a substantial underlying pattern within the data.

3.8 Correlation Analysis

A correlation matrix of the financial figures was examined to confirm the presence of significant correlations between variables. This examination supports the application of PCA and its ability to simplify the dataset by consolidating correlated variables into principal components.

3.9 Summary

The data preparation phase has successfully established a solid foundation for in-depth analysis. We have effectively cleaned, normalized, and distilled the data into its most informative elements. This preparation enables us to proceed with actuarial modeling and analysis, equipped to extract meaningful insights from the dataset.

4 Modelling

5 Evaluation

6 Deployment

Reference List

[1] *Actua Solutions*. es-ES. URL: https://actuasolutions.com (visited on 08/25/2023).

[2] Salem Al Gharbi et al. “Using Data-Mining CRISP-DM Methodology to Predict Drilling Troubles in Real-Time”. In: *Day 1 Tue, November 17, 2020*. Virtual: SPE, Nov. 2020, D013S103R013. DOI: 10.2118/202326-MS. URL: https://onepetro.org/SPEAPOG/proceedings/20APOG/1-20APOG/Virtual/451699 (visited on 08/25/2023).