Datathon 2025

Business problem

You are a junior data scientist at **FutureBright Insurance**. The underwriting department of the Automobile business line has requested your team's assistance in building a risk segmentation model using historical auto policy exposure and claims data. The goal of this project are threefold:

1. **Quantify expected losses** — and thus the risk level — for potential customers.
2. **Support the underwriting and actuarial departments** by providing insights into the nature of the business, such as:
   - Identify which segments of the portfolio are underpriced or overpriced.
   - Offer strategies to adjust rates accordingly.
3. **Assist the marketing department** in designing targeted campaigns to engage prospective customers.

Data

Your team has received two data sets for this initiative.

- `synthetic_auto_policies_model_data`: Contains exposure and claims information of 15000 historical policy terms.
- `synthetic_auto_policies_inference_data`: Contains information for 15000 future potential customers.

In the modeling dataset, key claim-related fields include:

- `claimcst0` - claim cost

- `clm` - claim indicator
- `numclaims` - claim counts

In the inference data, these fields are absent due to their unavailability for future customers.

For detailed definitions of each field, please refer to the Data Dictionary.

Modeling

After reviewing the data, your manager - a lead data scientist - decides to develop a quantified machine learning model to predict the claim cost per policy term (`claimcst0`). You are asked to follow standard modeling steps on this effort starting from model design, data collection and cleaning, data exploration, model selection and model implementation.

Final Delivery

Your final delivery should include two parts, with the focus on part 1.

1. Internal review (Technical Audience)
    - Target: Data scientists
    - Purpose: Describe the modeling procedures and techniques used at each stage, with particular emphasis on data exploration and model selection. Include both qualitative discussion and quantitative results from your analysis.
2. External review (Business Audience)
    - Target: Stakeholders (underwriters, actuaries, sales team)
    - Purpose: Address the original business questions in a clear and actionable way.

Note

The data used in this datathon is synthetic, derived from the dataset used in the Kaggle contest: 2023 UMN Travelers Analytics Case Competition.