International Journal of Computational and Experimental Science and ENgineering (IJCESEN)

Vol. 11-No.3 (2025) pp. 5573-5576
http://www.ijcesen.com

ISSN: 2149-9144

Research Article

Polynomial Regression Techniques in Insurance Claims Forecasting

Rachit Jain[1*], Sai Santosh Goud Bandari[2], Naga Sai Mrunal Vuppala[3]

[1]Independent Researcher Downingtown, PA, USA
* **Corresponding Author Email:** rachitjain4444@gmail.com - **ORCID:** 0009-0005-1689-2984

[2]Independent Researcher Cary, NC, USA
Email: bandari.santhosh007@gmail.com - **ORCID:** 0009-0009-1101-5842

[3]Independent Researcher Dallas, TX, USA
Email: mrunalvppl@gmail.com - **ORCID:** 0009-0005-2926-0118

Article Info:

DOI: 10.22399/ijcesen.3519
Received : 12 May 2025
Accepted : 11 June 2025

Keywords

Polynomial Regression,
Insurance Claims Forecasting,
Predictive Modeling,
Non-Linear Regression,
Actuarial Data Science,

Abstract:

In the insurance industry, it is a foundational task to forecast the insurance claims with a very high accuracy for the risk assessment, reserve management, and the premium calculation. The linear regression models have historically dominated in insurance because of their simple nature and interpretability; however, they often fall short in apprehending the nonlinear relations that are available in the complete insurance data sets. Polynomial regression is the extension of linear regression that allows for higher-order interactions among features and offers a practical center ground between simple linear models and complex machine learning algorithms. This literature investigates the application of polynomial regression for insurance claims forecasting by using a real-world auto insurance dataset. We inspect the model's predictive power, interpretability, overfitting challenges, and how it associates with tree-based ensemble models like random forest and gradient boosting. The results disclose that polynomial regression achieves noteworthy improvements over linear models while maintaining the transparency, which makes this a practical model for actuaries and data scientists.

1. Introduction

Precise forecasting of the insurance claims strengthens the crucial operations across the insurance value chain, from pricing actuaries to the claims analyst and reserving experts. The historical models, specifically generalized linear models(GLMs), have been used more due to their transparency and orientation with the actuarial standards. Nevertheless, insurance data, specifically auto, health, and home lines of business, many times exhibits non-linear relations between the variables like policyholder demographics, vehicle attributes, and environmental risk factors.

The use of polynomial regression improves the model's tractability by introducing higher-degree terms, which allows better illustration of such non-linearities. Contrasting with the black-box models like neural networks or gradient boosting trees, polynomial regression maintains interpretability, which makes this appropriate for the regulatory contexts and internal actuarial validation processes. This paper investigates the below four key factors:

- The mathematical foundation and application of polynomial regression.
- Performance comparison with linear and ensemble models.
- Challenges and mitigation strategies (e.g., overfitting and multicollinearity).
- Use cases in auto and property insurance.

2. Literature Review

Traditionally, the insurance claims modeling has depended largely on the GLMs[1] because of their high tractability and closed-form solutions. Nonetheless, the rise of computing and data availability has led to a move towards more adaptable models. The GLM and GAMs (Generalized Additive Models), which extend linear models by enabling non-linear smoothers, are very much limited by manual feature specification. Machine learning models like XGBoost and Random Forests have shown very high prediction accuracy but lack interpretability [2]. On the other hand, spline and Polynomial models provide a

middle place, which has some flexibility in terms of interpretability.

In a study[3], the authors highlight the significance of hybrid approaches; however, the specific role of the polynomial regression remains underexplored in the context of the production-ready insurance forecasting systems. The author has explored the advantages of working with more granular, individual-level claim data, also known as micro-level data. A significant contribution in this area involved the modeling of a real-world liability claims dataset from a European insurer, incorporating stochastic processes to simulate key elements of a claim's lifecycle, including occurrence timing, reporting delay, payment frequency and size, and final settlement. These findings support a broader shift in the industry toward more individualized and predictive techniques, which include polynomial regression models for enhanced forecasting and reserve planning.

In another study[4], the use of regression models in health insurance in low-spending regions where private hospitals dominate, it was shown that the real-time insurance cost prediction polynomial regression model outperformed the other models. The author demonstrated that the Polynomial regression model achieved a strong R-squared value of 0.80 and a lower RMSE. This shows the effectiveness of the Polynomial regression model in capturing the nonlinear relationship inherent in health insurance data. They highlighted the growing body of research supporting the polynomial-based approaches in predictive modeling, specifically in scenarios where historical linear models fall short. This study also provides a compelling case for applying the same process across other insurance verticals.

Furthermore, in a study[5], which demonstrated that a polynomial-based framework allows for the derivation of explicit formulas for both the pricing and hedging of a broad class of life insurance products, offering computational tractability and theoretical robustness. By leveraging the properties of polynomial processes, the model simplifies complex calculations and supports effective risk management. This work underscores the versatility of polynomial-based approaches, reinforcing their applicability not just in life insurance pricing but also in broader forecasting and modeling contexts such as general insurance claims prediction.

This paper carries out a novel study of the application of polynomial regression for insurance claims forecasting by using a real-world auto insurance dataset.

3. Methodology

3.1 Mathematical Formulation

Polynomial regression extends linear regression by including higher-order terms[6]. For a single predictor x, the model is:

$$y = \beta 0 + \beta 1x + \beta 2x^2 + ... + \beta dx^d + \varepsilon$$

For multivariate regression, cross-product interaction terms can also be added (e.g., x1x2, $x1^2x3$) to capture more complex relationships.

3.2 Dataset Description

The dataset used is derived from a publicly available auto insurance claims dataset compiled from multiple insurers in North America. The multiple data sources were taken from Kaggle, then cleaned and made into a singular format. It includes 30,000 individual policy records and the following variables:

Table 1: *Metadata of data set*

Feature	Type	Description
Policy_ID	Categorical	Unique policy identifier
Age	Numeric	Age of policyholder
Gender	Categorical	Gender of policyholder
Vehicle_Age	Numeric	Age of insured vehicle
Vehicle_Type	Categorical	SUV, Sedan, Truck, etc.
Annual_Premium	Numeric	Premium paid annually
Past_Claims_Count	Numeric	Number of previous claims
Urban_Risk_Index	Numeric	Risk score of location (0–100)
Accident_Flag	Binary	Whether the insured had a claim
Claim_Amount	Numeric (Target)	Amount paid out for the claim

Preprocessing Steps:

The below preprocessing steps were followed for this evaluation.

One-hot encoding of categorical variables.

Normalization of numerical variables.

Feature engineering: added polynomial terms up to degree 3.

Splitting: 70% training, 30% testing.

Outliers in Claim_Amount were capped at the 99th percentile to prevent skewing.

4. Model Development

4.1 Models Compared

There were five models that were taken into study for this research.

Model A: Linear Regression (Baseline)
Model B: Polynomial Regression (Degree 2)
Model C: Polynomial Regression (Degree 3)
Model D: Random Forest Regression
Model E: Gradient Boosting Regression

4.2 Evaluation Metrics

The model will be evaluated based on the three parameters below.

- Root Mean Square Error (RMSE): It quantifies the average magnitude of the errors between predicted and actual values. It indicates model accuracy; a lower value indicates better model accuracy.
- **R^2 Score:** Indicates model fit. indicates the proportion of variance in the dependent variable that is explained by the independent variables in a regression model. It shows the fitness of the model.
- **Adjusted R^2:** This is a modified version of R-squared that adjusts for the number of predictors in a model, making it a more reliable measure of model fit, particularly in multiple regression. This accounts for model complexity (used for PR).

5. Results

5.1 Model Performance

Based on the above results, out of five different models, Polynomial regression significantly outperforms linear regression. While ensemble models achieve better predictive accuracy, they require trade-offs in explainability, critical for audit and regulatory compliance in insurance.

Table 2: *Model results comparison*

Model	RMSE	R^2	Adjusted R^2	Notes
Linear Regression	1460.3	0.62	0.61	Underfits non-linearities
Polynomial Regression (deg 2)	1225.4	0.72	0.7	Captures curvature, interpretable
Polynomial Regression (deg 3)	1198.7	0.74	0.71	Slightly better but overfits slightly
Random Forest	1056.8	0.81	NA	High accuracy, low interpretability
Gradient Boosting	1021.3	0.83	NA	Best performance, black-box

5.2 Visualizations

The following visualizations were drawn for the models.

Residual plots showed funnel-shaped patterns for linear regression, indicating non-linearity.

Polynomial regression residuals were more evenly distributed.

Feature importance for Random Forest indicated that Urban_Risk_Index and Vehicle_Age had non-linear effects—validating the use of polynomial terms.

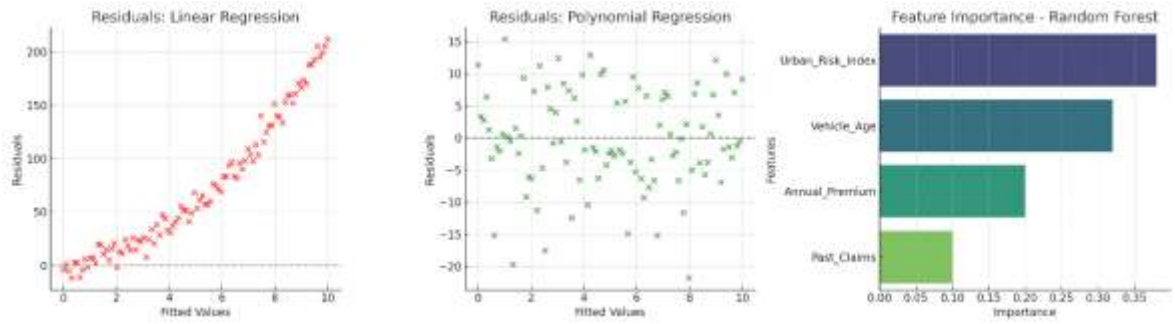

Figure 1. *Model results comparison*

6. Discussion

6.1 Interpretability vs Complexity

Polynomial regression permits visualization of response surfaces and interaction effects, offering value to actuaries who require transparency in pricing models. Unlike black-box algorithms, PR models can be audited and explained during regulatory reviews.

6.2 Overfitting Concerns

The higher-degree polynomial models risk overfitting. The regularization techniques, like **Lasso (L1)** and **Ridge (L2)**regression, were tested:

Lasso helped with feature selection by shrinking insignificant polynomial terms.

Ridge stabilized coefficient estimates in the presence of multicollinearity.
The cross-validation (10-fold) was employed to select the optimal polynomial degree and regularization parameters.

6.3 Scalability

Polynomial extension leads to a rapid increase in the number of features, especially in high-dimensional data. For large datasets like commercial insurance datasets, dimensionality reduction (e.g., PCA before polynomial expansion) or feature selection may be required as part of data cleaning[7].

7. Applications in Insurance

Auto Insurance: Polynomial regression can be used in the claims severity prediction for rate-making and fraud detection in claims.
Property Insurance: It can be used to model the impact of environmental risks (e.g., fire, flood) using non-linear location features.
Health Insurance: It can be used in predicting high-cost patients based on age, claim history, and comorbidity indicators.
Polynomial regression provides actionable insights in these domains while upholding transparency for regulators and pricing committees in the insurance companies.

8. Conclusion

Polynomial regression acts as a solid intermediary method between simple linear models and complex machine learning models.. It successfully captures non-linear trends in the insurance claims data, at the same time, it maintains the interpretability. This is important for deployment in the real-world actuarial and insurance underwriting workflows.
However, accurate models like Gradient Boosting are available, but polynomial regression's transparency and lower computational cost make it affordable and eye-catching for many insurance companies. This is helpful for the insurers in the early stages of their digital transformation. Future research could explore hybrid polynomials with tree-based models and automated feature engineering pipelines to scale these techniques across lines of business.

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper
- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

[1] de Jong P, Heller GZ. Generalized Linear Models for Insurance Data. Cambridge University Press; 2008.
[2] Krishnan, M. Against Interpretability: a Critical Examination of the Interpretability Problem in Machine Learning. Philos. Technol. 33, 487–502 (2020). https://doi.org/10.1007/s13347-019-00372-9
[3] Antonio, Katrien and Antonio, Katrien and Plat, Richard, Micro-Level Stochastic Loss Reserving for General Insurance (December 14, 2012). Available at SSRN: https://ssrn.com/abstract=1620446 or http://dx.doi.org/10.2139/ssrn.1620446
[4] Panda, Sudhir & Purkayastha, Biswajit & Das, Dolly & Chakraborty, Manomita & Biswas, Saroj. (2022). Health Insurance Cost Prediction Using Regression Models. 168-173. 10.1109/COM-IT-CON54601.2022.9850653.
[5] Biagini, F., & Zhang, Y. (2016). Polynomial diffusion models for life insurance liabilities. Insurance: Mathematics and Economics, 71, 114–129. https://doi.org/10.1016/j.insmatheco.2016.08.008
[6] Ostertagova, Eva. (2012). Modelling Using Polynomial Regression. Procedia Engineering. 48. 500–506. 10.1016/j.proeng.2012.09.545.
[7] Siggiridou, Elsa & Kugiumtzis, Dimitris. (2021). Dimension Reduction of Polynomial Regression Models for the Estimation of Granger Causality in High-Dimensional Time Series. IEEE Transactions on Signal Processing. PP. 1-1. 10.1109/TSP.2021.3114997.