COMP 9130 - Mini Project 1: Regression Challenge

Group 8

Mini Project 1 Report

Binger Yu
A01003660
gyu42@my.bcit.ca

Nicky Cheng
A01269051
ncheng25@my.bcit.ca

Supervisor: *Dr. Michal Aibin*

School of Computing and Academic Studies
Master of Science in Applied Computing
British Columbia Institute of Technology
Burnaby, British Columbia, Canada

Date: January 20, 2026

Contents

1 Introduction - Nicky

1.1 What problem are we solving?

We are trying to predict the cost of insurance based on BMI, Age, smoking status and region using regression models. We are also trying to see the best model for predicting health insurance costs.

1.2 Why is this problem important?

Accurately predicting medical insurance charges is important because healthcare costs can vary widely between individuals. A reliable regression model can help insurers and healthcare organizations estimate expected costs more consistently, support pricing and risk assessment, and improve financial planning for both providers and patients. From a machine learning perspective, this dataset also represents a realistic real-world regression problem that includes both numerical and categorical variables, as well as a right-skewed target distribution with outliers.

1.3 What is your target variable?

Our target variable is *charges*.

2 Data Exploration - Nicky

2.1 Dataset Statistics

- Samples: 1338
- Features: 7
- Types: 4 numerical, 3 categorical

2.2 Target Variable Distribution

Table I shows distribution of *charges*.

TABLE I
Distribution of variable *charges*.

Insurance Bracket	**Samples**
Under $10,000	712
$10,000 - $20,000	353
$20,000 - $30,000	111
$30,000 - $40,000	83
$40,000 - $50,000	52
$50,000 - $60,000	4
Over $60,000	3

2.3 Feature Correlations

Table II compares the correlations of each feature and *charges*.

TABLE II
Feature correlations to *charges*.

Feature	Correlation
smoker	0.79
age	0.30
bmi	0.20
children	0.07
region	0.006
sex	-0.06

2.4 Missing Values and Outliers

There are no missing values in the dataset. 5 values are extreme outliers in insurance charges. They have charges near or higher than $60,000.

3 Methodology - Binger

3.1 Data preprocessing steps

The preprocessing workflow included:

- **Train/test split:** 80/20 using `random_state=42`
- **Numeric preprocessing:** median imputation + StandardScaler
- **Categorical preprocessing:** most frequent imputation + OneHotEncoder

All preprocessing was implemented using a scikit-learn **Pipeline** and **ColumnTransformer** to ensure consistent transformations.

3.2 Feature engineering approach

Because `charges` is right-skewed, a target transformation experiment was included:

- Train using $\log(1 + \texttt{charges})$ (via `log1p()`)
- Convert predictions back using `expm1()`

3.3 Models selected and why

At least three models were trained and compared:

1. **Linear Regression** — baseline regression model.
2. **Ridge Regression** — L2 regularization to reduce overfitting and stabilize coefficients.
3. **Lasso Regression** — L1 regularization to encourage sparsity and reduce less useful coefficients.
4. **Ridge Regression (log target)** — to test whether a target transformation improves performance.

3.4 Evaluation metrics chosen

Models were evaluated using:

- **MAE** (Mean Absolute Error): interpretable average error magnitude.
- **RMSE** (Root Mean Squared Error): penalizes larger errors more strongly.
- R^2: measures how much variance is explained by the model.

4 Results - Binger

4.1 Model comparison table

Table III compares test performance for all models.

TABLE III
Model comparison results on the test set.

Model	**MAE**	**RMSE**	R^2
Linear Regression	4181.19	5796.28	0.7836
Lasso Regression	4181.19	5796.28	0.7836
Ridge Regression	4186.91	5798.30	0.7834
Ridge (log target)	3881.88	7780.62	0.6101

4.2 Best model performance

Based on **lowest RMSE**, the best-performing model was **Linear Regression**, with Ridge and Lasso producing very similar results.

4.3 Visualizations

The following visualizations were included in the modeling notebook:

- Predicted vs Actual plot (best model)

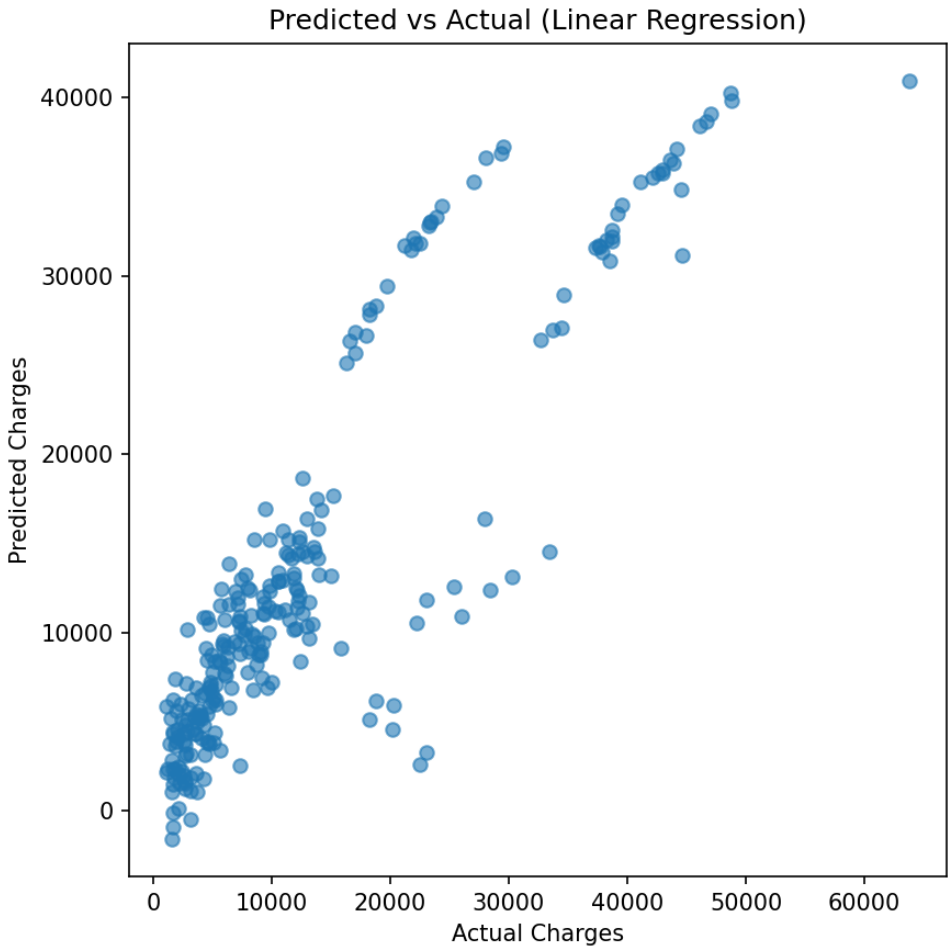

Fig. 1. Predicted vs Actual charges for the best-performing model on the test set.

- Residual plot (best model)

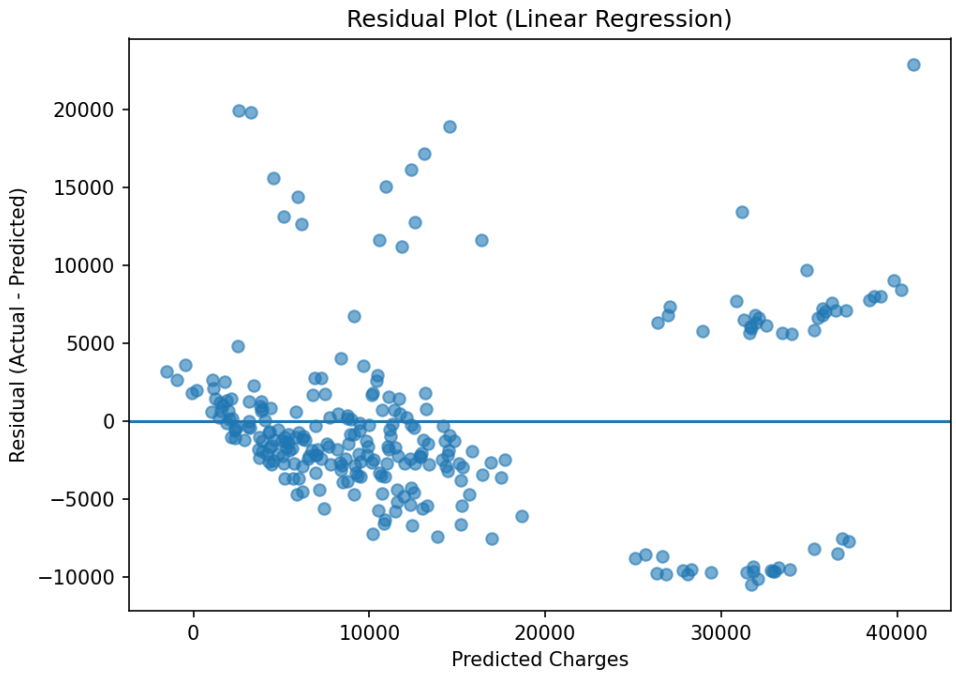

Fig. 2. Residual plot (Actual − Predicted) for the best-performing model.

4.4 Interpretation of results

Regularized linear models performed similarly to the baseline model. Errors increased for extreme high-charge cases, which is consistent with the right-skewed distribution of insurance charges.

5 Discussion - Binger

5.1 What worked well

- The preprocessing pipeline (ColumnTransformer + Pipeline) ensured numeric and categorical variables were processed consistently, preventing data leakage and keeping the workflow reproducible.
- Linear Regression, Ridge, and Lasso produced stable and very similar performance, showing that the relationship between features and charges can be captured reasonably well using linear models after proper encoding and scaling.
- Using multiple evaluation metrics (MAE, RMSE, and R^2) provided a more complete view of model quality, especially since the dataset contains high-charge outliers that affect RMSE strongly.

5.2 What did not work well

- Predictions were less accurate for extreme high-charge outliers, which is expected because insurance charges are right-skewed and contain a small number of very large values.
- The log-target Ridge model slightly improved MAE, but it produced a noticeably higher RMSE and lower R^2 compared with the baseline linear models.
- This trade-off happens because the log transformation compresses large charge values during training; however, after applying the inverse transform back to the original dollar

scale, a few large errors on high-charge cases can dominate RMSE since RMSE penalizes large errors more heavily than MAE.

5.3 What we would try next

- Use cross-validation and hyperparameter tuning to select the best regularization strength for Ridge and Lasso (e.g., tuning α).
- Experiment with non-linear models such as Random Forest Regression or Gradient Boosting to better capture complex interactions and improve performance on high-charge cases.
- Apply additional feature engineering, such as interaction terms (e.g., smoker $\times$ BMI) or polynomial features, since the EDA suggests smoking status has a strong relationship with charges.

5.4 Lessons learned

This mini project reinforced the importance of building end-to-end machine learning workflows that are reproducible and easy to evaluate. We learned that the choice of evaluation metric matters, especially when the target distribution contains outliers. Additionally, even simple regression baselines can perform strongly when preprocessing is done correctly, while feature transformations such as log-target scaling must be evaluated carefully because improvements in MAE may not translate to better RMSE or overall generalization.

6 Team Contributions - Binger

- **Nicky Cheng:** Completed `01_exploration.ipynb` (EDA, distributions, correlation checks, and initial written analysis).
- **Binger Yu:** Completed `02_modeling.ipynb` (preprocessing pipeline, feature engineering, training and evaluation of multiple regression models, results plots), finalized `requirements.txt` and `README.md`, corrected issues in `01_exploration.ipynb`, and prepared the Overleaf LaTeX report framework (template, formatting, and section titles).