Model Documentation and Results

Data description:

Features:

```
 DriverAge: Discrete
- CarAge: Discrete
- Brand: Qualitative
    -Renault, Nissan or Citroen            4
    -Japanese (except Nissan) or Korean    1
    -Opel, General Motors or Ford          3
    -Volkswagen, Audi, Skoda or Seat       5
    -Mercedes, Chrysler or BMW             2
    -Fiat                                  0
    -other                                 6
- Region: Qualitative
    -Centre                3
    -Ile-de-France         5
    -Bretagne              2
    -Pays-de-la-Loire      8
    -Aquitaine             0
    -Nord-Pas-de-Calais    7
    -Poitou-Charentes      9
    -Basse-Normandie       1
    -Haute-Normandie       4
    -Limousin              6
- Density: Continous
- Exposure: Continous
- Power: Qualitative
    -f    2
    -g    3
    -e    1
    -d    0
    -h    4
    -i    5
    -j    6
    -k    7
    -l    8
    -m    9
    -o    11
    -n    10
- Gas: Qualitative
    -Regular 1
    -Diesel  0
```

Response Variable:

```
ClaimNb: Discrete
```

Generalized Linear Model

Generalized linear models (GLMs) are a means of modeling the relationship between a variable whose outcome we wish to predict and one or more explanatory variables. The predicted variable is called the target variable and is denoted y. The predicted variable is called the target variable and is denoted y. For quantitative target variables, the GLM will produce an estimate of the expected value of the outcome. Here, we model the expected value of the number of claims of a policyholder. The mathematical specification of a GLM is as follows:

$$\begin{gathered} y_i \sim \text{Exponential}(\mu, \sigma^2) \\ g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} \\ \text{For the Poisson GLM,} \\ g(\mathbf{E}[Y]) = ln(\mathbf{E}[Y]) \end{gathered}$$

Why GLM?

Generalized Linear Models (GLMs) have been the workhorse of General Insurance pricing and reserving for the past thirty years. As linear models, they offer high interpretability through their coefficients and are both easy to use and understand. Additionally, GLMs meet actuaries' requirements by effectively modeling count events (such as claim frequency) and other non-normal response variables. In this project I use it as a benchmark to evaluate non-traditional models against.

Results:

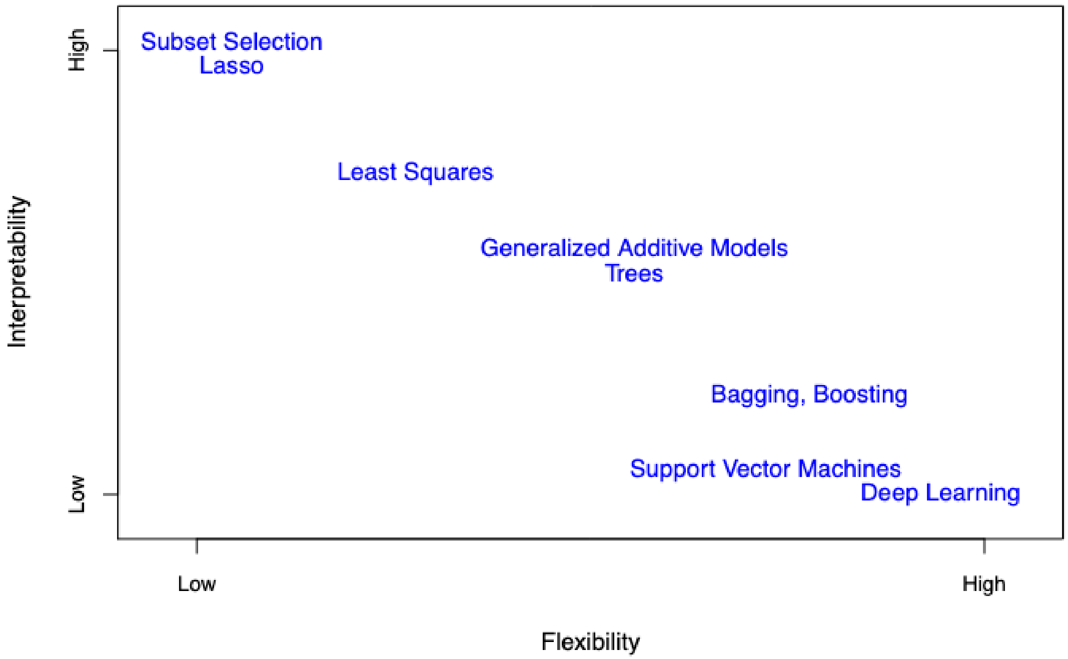
Model	Mean Poisson Deviance	Test MSE
GLM0 (homogeneous)	-	0.04609
GLM1 (all features)	0.26418	0.04577
GLM2 (without Brand)	0.26426	0.04578

Generalized Additive Models

The **generalized additive model (GAM)** is a GLM-like model that handles non-linearity natively. The mathematical specification of a GAM is as follows:

$$\begin{gathered} y_i \sim \text{Exponential}(\mu, \sigma^2) \\ g(\mu_i) = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + ... + f_p(x_{ip}) \\ \text{For the Poisson GAM,} \\ g(\mathbf{E}[Y]) = ln(\mathbf{E}[Y]) \end{gathered}$$

Why GAM?



GAMs are uniquely placed on the interpretability vs. predictive power continuum. In many applications they perform almost as well as more complex models, but are extremely interpretable.

- GAMs extend linear regression by allowing non-linear relationships between features and the target.
- The model is still additive, but link functions and multivariate splines facilitate a broad class of models.
- While GAMs are likely outperformed by non-additive models (e.g. boosted trees), GAMs are extremely interpretable.

Cubic Smoothing Splines

$$\text{Minimize} \sum_{i=1}^{n}(y - g(x_i))^2 + \lambda \int g''(t)^2 dt$$

where λ is the tuning parameter that controls flexibility. The splines that we are going to fit in our model are cubic as higher order splines lead to overfitting and/or higher standard errors.

Results:

Model	Test MSE
GAM (λ=100)	0.04593

The models were evaluated for different λ values using k-fold cross validation.

Conclusion

This project compared Generalized Linear Models (GLMs) and Generalized Additive Models (GAMs) for modeling claim frequency. The GLM with all features achieved the lowest test MSE of 0.04577, slightly outperforming the GAM (λ=100) with an MSE of 0.04593.

While GAMs provide flexibility in capturing non-linear relationships, GLMs remain highly effective and interpretable for insurance modeling. Future work could focus on optimizing GAM parameters or exploring advanced models for improved accuracy.