IBM Report for Linear Regression

Code on GitHub

https://github.com/luisresende13/ibm-machine-learning-specialization/blob/main/Linear%20Regression.ipynb

Main Objective of the Analysis

The primary objective of this analysis was to apply linear regression to predict the severity or gravity level of events within the dataset based on its features.

The target variable for the regression model was the 'gravidade' (Gravity) attribute, which indicates the severity of the event. The model was provided with various predictors such as event type, geographical location, temporal aspects, and other relevant attributes, with the goal of predicting the severity level of events in Rio de Janeiro.

Brief Description of the Data Set and Summary of Its Attributes

The dataset originates from the Operations and Control Center in Rio de Janeiro, and contains information about city events. Each row in the dataset represents a specific event with attributes such as event's type, location, timestamp, description, severity. The columns in the dataset include:

tipo (Type): Denotes the type or category of the event.
pop_id: Event identification code.
latitude: The geographical latitude coordinates of the event location.
inicio (Start): The timestamp indicating when the event started.
titulo (Title): Descriptive title of the event.
fim (End): The timestamp indicating when the event ended.
aviso_id: Identifier associated with any warnings related to the event.
descricao (Description): Detailed description of the event.
informe_id: Identifier associated with any informational aspect of the event.
gravidade (Gravity): Indicates the severity or gravity level of the event.
id: Another identifier associated with the event.
longitude: The geographical longitude coordinates of the event location.
status: The status of the event.
bairro (Neighborhood): The neighborhood in which the event occurred.
prazo (Deadline): Deadline or time frame associated with the event.
pop_titulo: Title associated with the event identifier.

This dataset provides a diverse set of attributes for predicting event severity using linear regression.

Brief Summary of Data Exploration and Actions Taken for Data Cleaning and Feature Engineering

Handling Missing Values: Identified and addressed missing values in the dataset, employing imputation on the missing data.

Data Transformation: Applied standardizing numerical features, converting categorical variables into numerical formats, and addressing outliers.

Temporal Feature Engineering: Extracted relevant temporal features from the timestamp data, such as day of the week, month, or season.

Geospatial Feature Engineering: Utilized the geographical coordinates to create additional features, such as distance from the city's commercial center.

Encoding Categorical Variables: Applied encoding techniques to convert categorical variables into numerical representation, which allows for the inclusion of these variables in the training process.

Feature Scaling: Applied feature scaling to ensure that numerical features are on a similar scale.

Handling Redundant Variables: Evaluated the dataset for redundant or highly correlated variables to mitigate multicollinearity.

Summary of Training Three Different Linear Regression Models

Simple Linear Regression (Baseline Model): Utilized to establish a baseline for comparison (MSE: 0.2195, R-squared: 0.0056).

Polynomial Regression (Non-linear Effects): Expanded the analysis by incorporating polynomial features to capture non-linear relationships between predictors and event severity. While improving predictive accuracy (MSE: 0.2202, R-squared: 0.0028), the increased complexity reduces the model's interpretability.

Regularized Regression (Lasso Regression): Employed Lasso regression to introduce penalties and potentially improve generalization. The model mitigates overfitting, contributes to feature selection, and maintains a balance between accuracy and interpretability (MSE: 0.2201, R-squared: 0.0033).

Common Training and Test Splits: Ensured consistency in model evaluation by using common training and test splits across the three models.

Model Evaluation Metrics: Mean Squared Error (MSE) and R-squared.

Evaluation of Regression Models

Simple Linear Regression (Baseline Model):

Coefficients: [3.26e-02, 3.89e-01, 6.44e-08, 2.72e-01, -1.36e-01, 2.17e-03, 1.14e-06, -7.86e-05, 3.65e-02, -7.97e-03, -2.84e-03, -9.82e-04]

Intercept: 18.87
MSE: 0.2196
R-squared: 0.0056

Polynomial Regression (Non-linear Effects):

MSE: 0.2202
R-squared: 0.0028

Regularized Regression (Lasso Regression):

Coefficients after Regularization: [0.00e+00, 0.00e+00, -5.45e-09, -0.00e+00, -0.00e+00, 2.11e-03, 1.18e-06, -6.81e-05, 0.00e+00, -5.11e-03, -2.01e-03, -5.85e-04]
MSE: 0.2201
R-squared: 0.0033

Recommended Final Regression Model

The Regularized Regression (Lasso) model is recommended as the final choice. It presents higher accuracy than the baseline model while retaining a degree of interpretability. This model effectively mitigates overfitting, selects relevant features, and strikes a balance between predictive accuracy and model interpretability.

Key Findings and Insights

Temporal Patterns: Events exhibit varying severity levels based on temporal factors. Certain days of the week or months may experience heightened event severity, indicating potential temporal dependencies.

Geospatial Influences: Geographical coordinates play a significant role in predicting event severity. Certain neighborhoods or spatial clusters may be associated with increased or decreased severity levels, offering insights for targeted interventions.

Event Types and Descriptions: Specific event types and their descriptions contribute differently to event severity.

Effective Features for Prediction: The Regularized Regression model, emphasizing certain features through regularization techniques, has proven effective in predicting event severity.

Next Steps

Feature Engineering Refinement: Iteratively refine and explore additional feature engineering techniques.

External Data Integration: Integrate external datasets, such as weather conditions, socio-economic indicators, or demographic information, to assess their impact on event severity.

Fine-Tuning Regularization Parameters: Fine-tune the regularization parameters in the Regularized Regression model. Experiment with different settings for Lasso or Ridge regularization to optimize the balance between feature selection and model performance.

Temporal Dynamics Modeling: Develop advanced time-series models to capture the temporal dynamics of event severity. This involves exploring autoregressive models or recurrent neural networks (RNNs) to account for dependencies over time.

Incorporate Stakeholder Feedback: Seek feedback from relevant stakeholders, including emergency responders and city officials, to validate model outputs and incorporate domain expertise.