IBM Report for Classification Project

Code on GitHub

https://github.com/luisresende13/ibm-machine-learning-specialization/blob/main/Classification.ipynb

Main Objective of the Analysis

The main goal of this analysis was to employ a classification model to predict the categorical severity levels of events within the dataset based on its features.

The target variable for the classification model was the 'gravidade' (Gravity) attribute, which signifies the severity of the event. The model utilized diverse predictors such as event type, geographical location, temporal aspects, and other relevant attributes. The aim was to categorize events in Rio de Janeiro into distinct severity levels.

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 classification.

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 Classification Models

Logistic Regression: Utilized for baseline comparison, the Logistic Regression model achieved an accuracy of 35.71%. It demonstrated precision, recall, and F1-score (weighted averages) of 43%, 36%, and 36%, respectively.

Decision Tree: The Decision Tree model, also with an accuracy of 42.86%, displayed precision, recall, and F1-score (weighted averages) of 57%, 43%, and 44%, respectively.

Random Forest: Outperforming others, the Random Forest model achieved an accuracy of 47.62%. It exhibited precision, recall, and F1-score (weighted averages) of 60%, 48%, and 49%, respectively.

In addition to accuracy, precision, recall, and F1-score, confusion matrices were used to evaluate and compare the performance of each classification model.

Results of Classification Models

<u>Logistic Regression</u>:

Classification Report:

	precision	recall	f1-score	support
-1	0.80	0.44	0.57	9
0	0.00	0.00	0.00	5

1	0.50	0.38	0.43	8
2	0.44	0.50	0.47	8
3	0.33	0.14	0.20	7
4	0.21	0.60	0.32	5
accuracy			0.36	42

Confusion Matrix:

```
[[4 0 0 0 0 5]
[0 0 1 3 0 1]
[1 0 3 1 1 2]
[0 4 0 4 0 0]
[0 1 1 1 1 3]
[0 0 1 0 1 3]]
```

Decision Tree:

Classification Report:

	precision	recall	f1-score	support
-1	1.00	0.33	0.50	9
0	0.25	0.40	0.31	5
1	0.40	0.25	0.31	8
2	0.83	0.62	0.71	8
3	0.33	0.29	0.31	7
4	0.29	0.80	0.42	5
accuracy			0.43	42

Confusion Matrix:

```
[[3 0 0 0 0 6]
[0 2 0 1 1 1]
[0 2 2 0 3 1]
[0 2 0 5 0 1]
[0 2 2 0 2 1]
[0 0 1 0 0 4]]
```

Random Forest:

Classification Report:

	precision	recall	f1-score	support

-1	0.75	0.33	0.46	9
0	0.27	0.60	0.37	5
1	0.60	0.38	0.46	8
2	0.67	0.50	0.57	8
3	0.75	0.43	0.55	7
4	0.33	0.80	0.47	5
accuracy			0.48	42

Confusion Matrix:

```
[[3 0 0 0 0 6]
[0 3 0 2 0 0]
[1 2 3 0 1 1]
[0 3 0 4 0 1]
[0 3 1 0 3 0]
[0 0 1 0 0 4]]
```

Recommended Final Regression Model

The Random Forest was chosen as the recommended final model. It achieved the highest accuracy among the models, reaching 47.62%. The precision, recall, and F1-score (weighted averages) of 60%, 48%, and 50%, respectively, highlight its superior performance in predicting the severity or gravity level of events. The Random Forest model demonstrates a balanced trade-off between precision and recall, making it a good choice for the classification task.

Key Findings and Insights

The three classification models, Logistic Regression, Decision Tree, and Random Forest, resulted in similar accuracy of 35.71%, 42.86% and 47.62% respectively. However, the Random Forest model outperformed others with the highest accuracy (47.62%), indicating its effectiveness in classification tasks.

Precision-Recall Trade-off: The model demonstrated a balanced performance with precision, recall, and F1-score. This suggests that the Random Forest algorithm effectively captures the complexity and patterns within the dataset.

Temporal Patterns: Feature importance analysis revealed that 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: It was also demonstrated that 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.

The classification task faced challenges, especially with class imbalance, where certain severity levels had fewer instances. This imbalance may impact the model's ability to generalize well for specific categories.

Next Steps

Hyperparameter Tuning: Further optimization of model hyperparameters may lead to improvements in overall performance.

Feature Engineering Refinement: Exploring additional features or refining existing ones could potentially increase the predictive power of the models.

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

Implementation Considerations: When deploying the recommended Random Forest model, it is crucial to monitor its performance in real-world scenarios and adapt as necessary.

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