Dear Group 24,
Thank you for submitting your notebooks. Below, I have provided a detailed review of your work, highlighting both strengths and areas for improvement.
Your submission demonstrates significant effort and the application of multiple data science techniques. However, there are areas where methodological rigor, clarity, and execution could be improved.
You submitted two notebooks, each running from start to finish, along with three helper files. While the work is functional, certain methodological and interpretational shortcomings need attention.

1. Imports & Data Exploration

Imports:

- All necessary packages and files are imported from the beginning, ensuring seamless integration.

Descriptive Statistics:

- Main descriptive statistics are reviewed with sections guided by ChatGPT-generated descriptions. However, interpretation is limited, and critical steps, like checking for duplicate identifiers, are missing.
- The target distribution is eventually checked through visualization, which should ideally occur earlier in the workflow.

Incoherences:

- Key inconsistencies are identified, including minors being injured, small Years of Birth, and columns with full NaNs.
- Outlier treatment (e.g., IME-Count) is applied but treated as an inconsistency, which may require a clearer explanation.
- Some NaN zip codes are improperly handled. Full NaN columns are dropped appropriately.

Visual Exploration:

- Sparse univariate distribution charts and boxplots for outliers are provided but lack any interpretation. Insights from these visualizations would add significant value.

Multivariate Relationships:

- Categorical bar charts with hue are included, but again, no interpretation is provided, limiting their usefulness.

2. Preprocessing & Cleaning

Missing Values:

- Medians are used for date variables, and 0 is filled for codes and dates. However, using 99999 as a placeholder for ZipCode may cause scaling issues later.
- KNN imputation is applied before scaling, which may lead to inconsistencies.

Outliers:

- Robust Scaler is attempted.
- IME-4 Count removal is applied early in the workflow but requires justification.

Categoricals:

- Frequency Encoding is applied before separating the data, causing leakage. Counts should be based only on the training set. Binary Encoding also suffers from the same issue, even if it works for this dataset.

- Ordinal Encoding is correctly implemented. Dummy Encoding is used but without dropping the first category. In this case, OHE would be better than get_dummies.

3. Feature Engineering & Selection

Feature Engineering:

- Date variables and binary flags are created. However, more advanced numerical transformations or date differences could add value.

Data Scaling:

- Separate scalers (Standard, MinMax, and Robust) are tested.
- Observations made about the effects of scalers (e.g., skewness in Birth Year) suggest some misunderstanding of their role, as scalers do not inherently adjust data distributions. Subsequent steps primarily use Robust Scaled variables.

Feature Selection:

- Variance Threshold (set at 0.1) is applied unnecessarily as all variances exceed this threshold.
- Chi-Squared is applied incorrectly, treating Frequency Encoded variables as numerical.
- Recursive Feature Elimination (RFE) is performed, but results suggest it is assessed on training data, which can lead to overfitting.
- The justification for feature selection is unclear, particularly regarding the inclusion/exclusion of categorical variables in ExtraTrees, RFE, and Lasso models.

4. Modeling & Evaluation

Model Training and Implementation:

- Multiple models are trained, including Logistic Regression, SVM with SGD, Decision Trees, Random Forest, Gradient Boosting, AdaBoost, LDA (using SVD), QDA, and XGBoost.
- Grid Search is applied to Random Forest, which performs best, though Gradient Boosting achieves similar results.
- Advanced techniques like Stacking, Voting, and Calibrated SGD classifiers are attempted, but these may add unnecessary complexity. Stick to models you are comfortable explaining in-depth.

Model Assessment:

- F1-Macro is used as the primary metric, which is appropriate. However, additional metrics should be included for a more comprehensive evaluation.
- Avoid using the Kaggle score as the primary metric for judgment. Instead, rely on validation scores at this stage.

Cross-Validation:

- Cross-validation is attempted but ultimately abandoned. Proper stratified CV should be implemented to improve reliability.

5. PDF

Intended Structure:

- The report is a detailed account of what was done and generally follows the intended structure, albeit non-intentionally.

Quality:

- The report is overly detailed in some areas, mentioning unnecessary steps like dropping the target column.
- Results are presented in lengthy blocks of text, and screenshots of tables detract from professionalism.

Level of Detail:

- Variables, transformations, and models are mentioned but lack specificity and insights, particularly regarding the rationale for key decisions.

Open-Ended Section:

- The inclusion of an analytics interface via Streamlit is a nice touch, with some detail provided on its functionality.

Final Recommendations

1. **Visual Exploration**: Add interpretations for visualizations to derive meaningful insights.
2. **Preprocessing**: Avoid data leakage by ensuring preprocessing (e.g., frequency encoding and scaling) is applied only to the training data.
3. **Feature Selection**: Refine feature selection techniques and provide stronger justifications for decisions.
4. **Modeling**: Simplify the approach to focus on models you can explain confidently. Expand evaluation to include additional metrics for a holistic analysis.
5. **Cross-Validation**: Implement proper stratified CV to evaluate models without leakage.
6. **Report Writing**: Streamline the report to focus on key findings and justifications rather than lengthy procedural details.
7. **Open-Ended Section**: Further develop the Streamlit analytics interface and provide more depth in its presentation.

Your work demonstrates effort and potential. Addressing these issues will significantly improve the rigor and clarity of your submission. Please let me know if you have any questions or need further assistance.

Best regards,

Ricardo Santos