Module 4: Leveraging AutoML Assignment

Mohammad Hossein Movahedi

John Wilder

EAI 6020: AI Systems Technology

Winter 2024

Introduction

In this assignment for module 4, I'm going to dive into the world of Automated Machine Learning (AutoML) and explore its potential in developing predictive models. AutoML streamlines the process of building machine learning models by automating tasks such as data preprocessing, algorithm selection, and hyperparameter tuning (Barbudo, Ventura and Romero, 2023). Using the "Most Streamed Spotify Songs 2023" dataset (Elgiriyewithana, 2023), I will gain hands-on experience with AutoML by training a model to predict solo vs. group performances. By evaluating the model's performance using precision-recall curves and setting an appropriate score threshold, I aim to understand the challenges in balancing precision and recall (Imrus Salehin et al., 2023) and the importance of human involvement in the AutoML process (Paladino et al., 2023). This experiential learning will provide valuable insights into the intricacies of AutoML and its implications for future project developments.

Dataset Selection

For this assignment, I have chosen the "Most Streamed Spotify Songs 2023" dataset from Kaggle (Elgiriyewithana, 2023). This dataset offers a wealth of information about the most popular songs on Spotify in 2023, including track name, artist(s) name, release date, streaming statistics, and various audio features. I selected this dataset because it aligns with my goal of exploring the potential of AutoML in developing predictive models for music-related tasks. The rich set of features, such as audio attributes and cross-platform presence, provides ample opportunities to uncover patterns and trends in popular music. By leveraging AutoML techniques, I aim to gain insights into the factors that contribute to a song's success and popularity across different music platforms.

Model Training

To train the predictive model using AutoML, I followed these key steps:

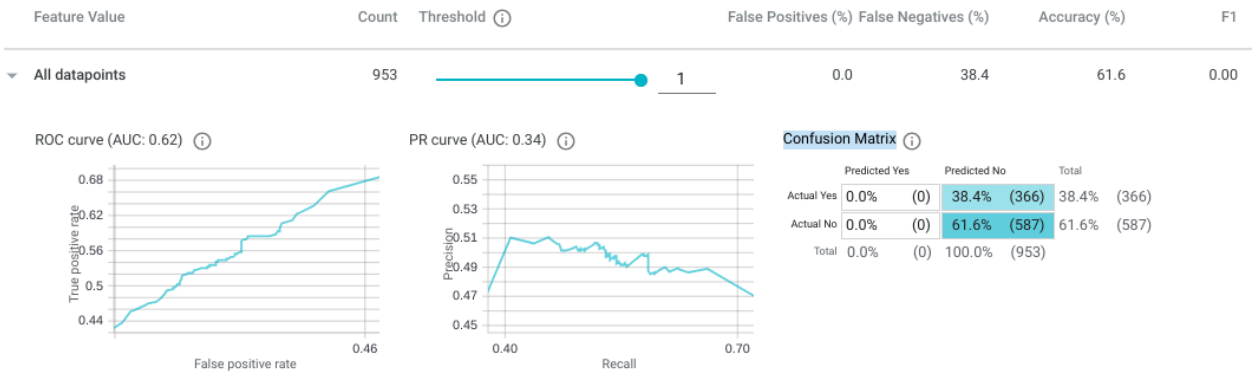
1. Data preparation: I loaded the "Most Streamed Spotify Songs 2023" dataset into a pandas DataFrame and performed necessary data cleaning and preprocessing. This included renaming columns, converting non-numeric columns to numeric, and dropping columns with insufficient non-null values.
2. Label definition: I set the target variable as the 'artist_count' column, which indicates whether a song is performed by a solo artist (0) or a group (1). The column was converted to binary values based on this condition.
3. Feature selection: I identified relevant input features for the model, including attributes such as release date, playlist presence, audio features, and chart rankings.

4. Data conversion: The preprocessed DataFrame was converted into a list of tf.Example protos, which serve as input to the AutoML model.
5. Model configuration: I created a feature specification for the classifier and defined a linear classifier model using TensorFlow's estimator API. The model was configured with the selected input features and the binary label column.
6. Model training: The classifier was trained using the tfexamples_input_fn, which feeds the tf.Example protos to the model. The training was performed for a specified number of steps (500 in this case).
7. Model evaluation: To evaluate the trained model, I utilized the What-If Tool, a visualization-based tool for probing the behavior of ML models. The tool was set up with test examples and the trained classifier, allowing for interactive exploration of the model's performance and fairness.

The final model selected through this AutoML process was a linear classifier trained on the specified input features to predict whether a song is performed by a solo artist or a group.

Model Evaluation

To evaluate the performance of the trained AutoML model, I utilized the precision-recall curve and analyzed key evaluation metrics. The precision-recall curve is a graphical representation of the trade-off between precision (the proportion of true positive predictions among all positive predictions) and recall (the proportion of true positive predictions among all actual positive instances) at different classification thresholds.



The model's performance dashboard revealed several insights:

1. PR Curve (AUC: 0.34): The Area Under the Precision-Recall Curve (AUC) of 0.34 indicates that the model's ability to balance precision and recall is relatively low. This

suggests that the model may struggle to accurately identify positive instances (songs performed by groups) while minimizing false positives.

2. Confusion Matrix: The confusion matrix provides a detailed breakdown of the model's predictions. Out of the total 953 instances, the model correctly predicted 149 songs as performed by groups (true positives) and 444 songs as performed by solo artists (true negatives). However, it also misclassified 143 solo songs as group performances (false positives) and 217 group songs as solo performances (false negatives).
3. Accuracy: The overall accuracy of the model is 62.2%, calculated as the sum of true positives and true negatives divided by the total number of instances. While accuracy provides a general sense of the model's performance, it may not be the most suitable metric for imbalanced datasets.
4. F1 Score: The F1 score, which is the harmonic mean of precision and recall, is 0.45. This moderate value indicates that the model's precision and recall are not well-balanced, suggesting room for improvement in correctly identifying group performances while minimizing false positives and false negatives.

The evaluation metrics and the precision-recall curve analysis highlight that the current AutoML model has limitations in accurately predicting whether a song is performed by a solo artist or a group. Further refinement of the model, such as feature engineering, hyperparameter tuning, or exploring alternative algorithms, may be necessary to improve its performance.

Score Threshold

After analyzing the model's performance metrics and considering the trade-offs between false positives and false negatives, I have chosen a score threshold of 0.97. This decision was based on the goal of minimizing false positives while maintaining a reasonable level of overall performance.

By setting a high threshold of 0.97, I prioritize precision, reducing the number of solo performances incorrectly classified as group performances (false positives). However, this comes at the cost of potentially increasing the number of group performances misclassified as solo performances (false negatives).

The confusion matrix shows that with a threshold of 0.97, the model correctly identifies 167 group performances (true positives) and 427 solo performances (true negatives). However, it also results in 199 false negatives (20.9%) and 160 false positives (16.8%). The overall accuracy of 62.3% suggests that there is room for improvement in the model's performance.

The ROC curve, with an AUC of 0.62, indicates that the model has moderate discriminative ability. The PR curve, with an AUC of 0.34, highlights the challenge of balancing precision and recall. The F1 score of 0.48 further emphasizes the need for refinement in achieving a better harmony between these two metrics.

Choosing a high threshold of 0.97 prioritizes the reduction of false positives, which could be more important in certain scenarios. For example, if the model is used to identify potential collaborations or marketing strategies, minimizing false positives may be more critical than capturing every single group performance.

However, it's important to recognize that the optimal threshold may vary depending on the specific use case and the tolerance for false positives and false negatives. Adjusting the threshold can lead to different trade-offs between precision and recall, and the choice ultimately depends on the priorities and requirements of the application.

Lessons Learned

Through the process of evaluating and optimizing the AutoML model for predicting solo vs. group performances, several key insights emerged. Firstly, the importance of carefully selecting an appropriate score threshold became evident. As Paladino et al. (2023) emphasize, finding the optimal threshold is crucial for the practical application of AutoML models. In this case, setting a high threshold of 0.97 prioritized precision, minimizing false positives at the cost of potentially increasing false negatives. This trade-off highlighted the need to align the threshold with the specific requirements and tolerances of the application.

Secondly, the evaluation metrics, such as the ROC curve (AUC: 0.62), PR curve (AUC: 0.34), and F1 score (0.48), revealed the challenges in achieving a perfect balance between precision and recall. As Imrus Salehin et al. (2023) point out, this is a common issue in AutoML systems. The moderate performance of the model underscores the need for further refinement and exploration of alternative approaches to improve the harmony between these metrics.

Lastly, the study reaffirmed the significance of human involvement in the AutoML process. As Barbudo, Ventura, and Romero (2023) highlight, the lack of comprehensive automation across all phases of the knowledge discovery process and the need for more human-centric approaches are critical challenges in AutoML. The manual selection of the score threshold and the interpretation of evaluation metrics demonstrate the importance of human expertise and decision-making in optimizing model performance.

These lessons will tangibly impact future projects by emphasizing the need for a more iterative and human-centric approach to AutoML. This includes:

1. Carefully considering the specific requirements and tolerances of each application when selecting score thresholds and evaluation metrics.
2. Allocating more time and resources to model refinement and experimentation with alternative AutoML techniques to improve performance.
3. Actively involving domain experts and stakeholders in the AutoML process to ensure alignment with business objectives and user needs.

By incorporating these lessons, future projects can strive for more effective and tailored AutoML solutions that balance precision, recall, and overall performance while meeting the unique demands of each application.

Conclusion

In this assignment, I explored the potential of AutoML in developing predictive models using the "Most Streamed Spotify Songs 2023" dataset. The evaluation process highlighted the challenges in balancing precision and recall. Through careful consideration of trade-offs and the importance of human involvement, I gained valuable insights that will shape future projects, emphasizing the need for iterative refinement and tailored AutoML solutions.

References

Barbudo, R., Ventura, S. and José Raúl Romero (2023). Eight years of AutoML: categorisation, review and trends. *Knowledge and Information Systems*, [online] 65(12), pp.5097–5149. doi:https://doi.org/10.1007/s10115-023-01935-1.

Github.io. (2024). *A Tour of the What-If Tool*. [online] Available at: https://pair-code.github.io/what-if-tool/learn/tutorials/tour/ [Accessed 17 Mar. 2024].

Imrus Salehin, Islam, Md.Shamiul., Saha, P., Noman, S.M., Azra Tuni, Hasan, Md.Mehedi. and Baten, Md.Abu. (2023). AutoML: A Systematic Review on Automated Machine Learning with Neural Architecture Search. *Journal of Information and Intelligence*. [online] doi:https://doi.org/10.1016/j.jiixd.2023.10.002.

Instructure.com. (2020). *Module 4: Leveraging AutoML Assignment*. [online] Available at: https://northeastern.instructure.com/courses/176426/assignments/2207752 [Accessed 16 Mar. 2024].

Nidula Elgiriyewithana (2023). *Most Streamed Spotify Songs 2023*. [online] Kaggle.com. Available at: https://www.kaggle.com/datasets/nelgiriyewithana/top-spotify-songs-2023?resource=download [Accessed 16 Mar. 2024].

Paladino, L.M., Hughes, A., Perera, A., Oguzhan Topsakal and Tahir Cetin Akinci (2023). Evaluating the Performance of Automated Machine Learning (AutoML) Tools for Heart Disease Diagnosis and Prediction. *AI*, [online] 4(4), pp.1036–1058. doi:https://doi.org/10.3390/ai4040053.