INFO7390: Advances Data Sci/Architecture

Project 2: Supervised learning with feature engineering

Naman Gupta
(NUID: 002729751)
Northeastern University
Boston Campus
gupta.naman@northeastern.edu

Introduction

Welcome to a journey through supervised learning and feature engineering, all in the context of a classification problem. We started with a crystal-clear understanding of our goals and target variables. Our adventure took us through the collection and preparation of data, where we carefully curated the dataset. We then entered the realm of feature engineering, enhancing our model's predictive power. With data ready, we selected the right algorithm using cross-validation and trained our model. Model evaluation was the final test, revealing strengths and weaknesses.

This report offers concise documentation of our process, findings, and recommendations for future work. While ethical considerations are important, we also reflect on our learning journey in the dynamic field of machine learning.

Problem Statement

The dataset has been acquired from the "UC Irvine Machine Learning Repository" and can be accessed via the following link: https://archive.ics.uci.edu/dataset/222/bank+marketing

This dataset revolves around a Portuguese banking institution's direct marketing campaigns, specifically focusing on phone calls made to clients. The primary goal of this dataset is to predict whether a client will opt to subscribe to a term deposit, a binary outcome indicated by the variable "y." These marketing endeavors typically entailed multiple interactions with each client to ascertain whether they eventually decided to subscribe to the bank's term deposit product, which is recorded as either 'yes' or 'no.'

Our primary classification task centers around forecasting whether a client will subscribe to the term deposit (yes or no).

Our dataset revolves around direct marketing efforts involving phone calls made by a Portuguese banking institution. The primary objective is to predict whether a client will agree to subscribe to a term deposit, marked as 'yes' or 'no' in the variable "y."

Here's a summary of the input variables:

- Age (numeric)
- Job (categorical): Type of job
- Marital (categorical): Marital status

- Education (categorical)
- Default (binary): Credit in default?
- Balance (numeric): Average yearly balance in euros
- Housing (binary): Has a housing loan?
- Loan (binary): Has a personal loan?
- Contact (categorical): Communication type
- Day (numeric): Last contact day of the month
- Month (categorical): Last contact month of the year
- Duration (numeric): Last contact duration in seconds
- Campaign (numeric): Number of contacts performed during this campaign and for this client
- Pdays (numeric): Number of days since the client was last contacted from a previous campaign (-1 means not previously contacted)
- Previous (numeric): Number of contacts performed before this campaign and for this client
- Poutcome (categorical): Outcome of the previous marketing campaign
- Y (binary): Has the client subscribed to a term deposit (yes or no)?

Solution

Understanding and Definition of the Problem:

The solution development commences with an initial step in predictive modeling: Exploratory Data Analysis, including y-data profiling. This analysis yields essential insights into the dataset, encompassing the distribution of data and the correlations between features. Some key observations from the y-data profiling include the presence of 7 numerical features, 6 categorical features (including the target variable), and 4 boolean features. The dataset exhibits exceptional data quality, with no missing values or duplicates. Notably, the client's age follows a normal distribution with an average age of 40. Additionally, most clients are married, have secondary education, no personal loans, and were contacted via cellular phones. There is a strong positive correlation between the number of contacts performed before the current campaign and the number of days passed since the client's last contact in the previous campaign. This suggests that clients who were not contacted earlier for an extended period were approached more frequently. Key features such as the duration of the last contact, the outcome of the previous campaign, and the last month of contact are highly related to the success of the client's subscription to a term deposit. While all features exhibit positive correlations with the target variable, none of them significantly dominate in predicting the target variable.

`pdays` is highly overall correlated with `previous` and 1 other fields	High correlation
`previous` is highly overall correlated with `pdays`	High correlation
`housing` is highly overall correlated with `month`	High correlation
`contact` is highly overall correlated with `month`	High correlation
`month` is highly overall correlated with `housing` and 1 other fields	High correlation
`poutcome` is highly overall correlated with `pdays`	High correlation
`default` is highly imbalanced (87.0%)	Imbalance
`poutcome` is highly imbalanced (53.1%)	Imbalance
`previous` is highly skewed (γ1 = 41.84645447)	Skewed
`balance` has 3514 (7.8%) zeros	Zeros
`previous` has 36954 (81.7%) zeros	Zeros

Figure 1: Additional insights from y-data profiling

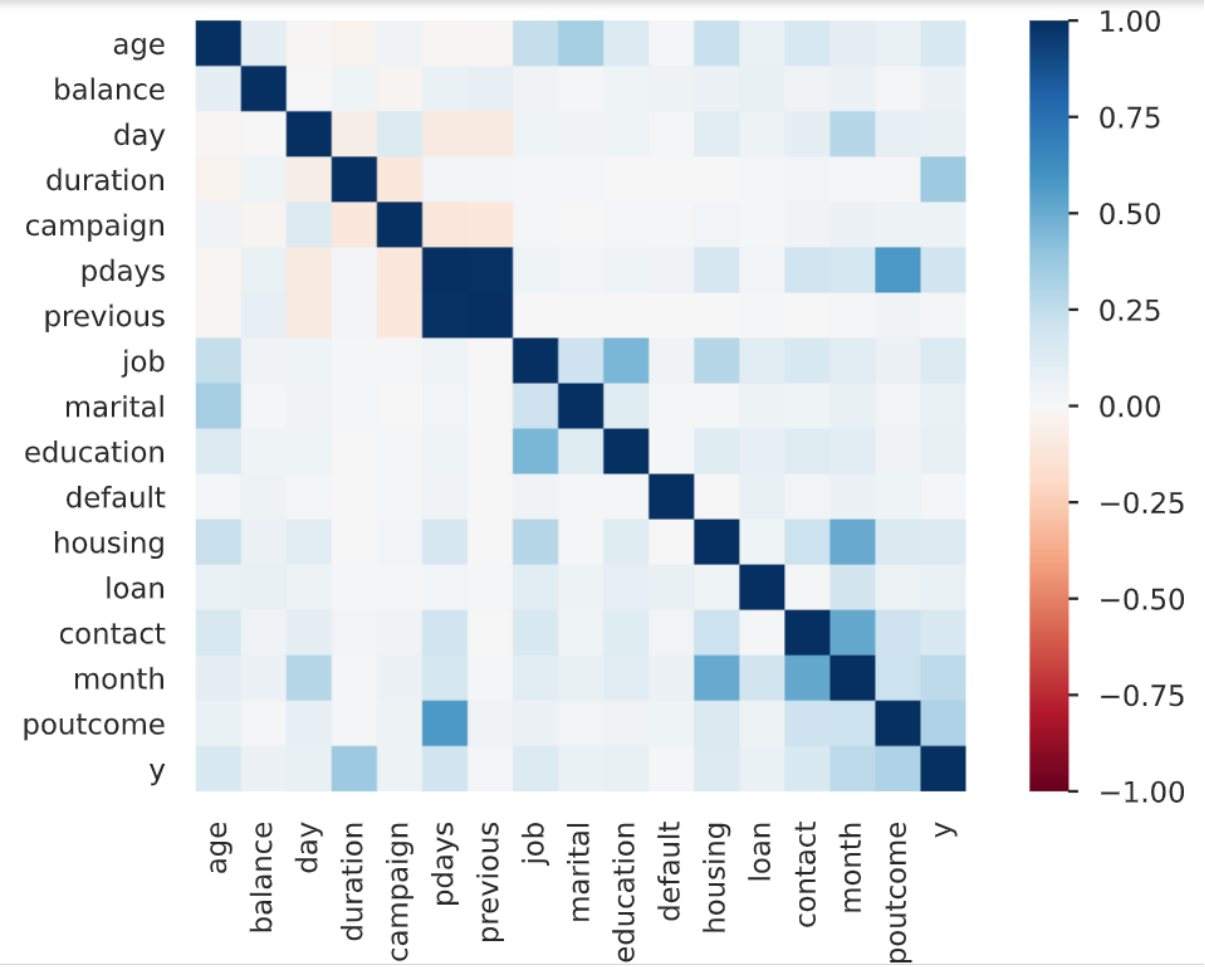

Figure 2: Heat Map for feature correlation

Data Collection and Preparation:

Upon gaining a comprehensive understanding of the data, it becomes evident that there are no missing values or outliers in the dataset, signifying its quality. Subsequently, we examine the datatypes of all features and proceed to split the data into training and testing sets, adhering to an 80:20 ratio. This partitioning allocates 80% of the data for training the model and reserves the remaining 20% for evaluating the model's performance.

Feature Engineering:

In the realm of feature engineering, we introduce a novel feature called "date_bucket" to replace the literal day of the month. This categorization into 'Early Month,' 'Mid Month,' 'Late Month,' or 'End of Month' adds value to predictive modeling. It addresses the issue of managing 31 separate day categories. Further examination reveals that categorical columns exhibit distinct values ranging from 2 to 12, without a natural order. To handle these features, we employ one-hot encoding due to their multiple classifications. In contrast, label encoding is reserved for the target variable. For numerical features, min-max scaling is applied to bring them to a common scale, mitigating any disproportionate influence due to varying magnitudes.

	age	balance	duration	campaign	pdays	previous	job_admin.	job_blue-collar	job_entrepreneur	job_housemaid	...	month_jun	month_mar	m
0	0.298701	0.080511	0.014640	0.000000	0.0	0.0	0.0	1.0	0.0	0.0	...	0.0	0.0	
1	0.402597	0.085650	0.054697	0.016129	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	
2	0.311688	0.107684	0.026434	0.048387	0.0	0.0	1.0	0.0	0.0	0.0	...	0.0	0.0	
3	0.246753	0.071723	0.076251	0.161290	0.0	0.0	0.0	0.0	0.0	0.0	...	1.0	0.0	
4	0.493506	0.104561	0.053680	0.016129	0.0	0.0	0.0	1.0	0.0	0.0	...	0.0	0.0	

Figure 3: Dataframe after preprocessing and feature engineering

Model Selection and Development:

The selection and development of the predictive model involve considering several potential classification models, including Random Forest, Gradient Boosting, SVM, Logistic Regression, and Decision Tree. The decision regarding the optimal model hinges on cross-validation, employing 5-fold validation and accuracy as the evaluation metric. Gradient Boosting emerges as the top-performing model with an accuracy score of 0.91 on the training dataset and 0.89 on the testing dataset.

Model Evaluation:

To assess the model's performance, we delve into feature importance analysis, revealing that the duration of the last contact with the client and the outcome of the previous campaign play pivotal roles in determining whether the client subscribes to the service post-campaign. Evaluating the model further, we examine the confusion matrix, identifying 7424 true positives. Additionally, the Receiver Operating Characteristic (ROC) curve analysis showcases the model's proficiency. It surpasses the random classifier, with an area under the curve (AUC) of 0.91, indicating a highly effective classifier. In summary, the model demonstrates exceptional performance, aligning with the objectives of the problem statement.

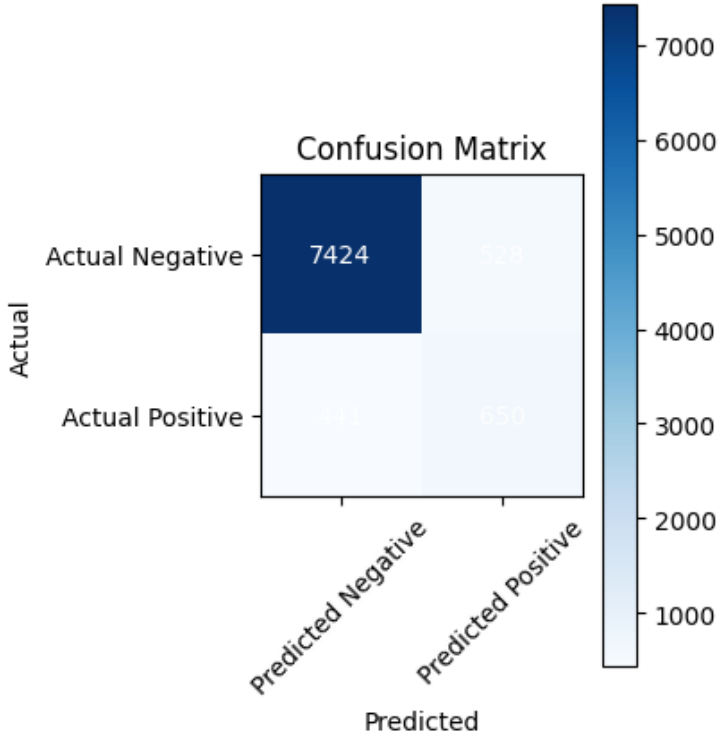

Figure 4: Confusion Matrix

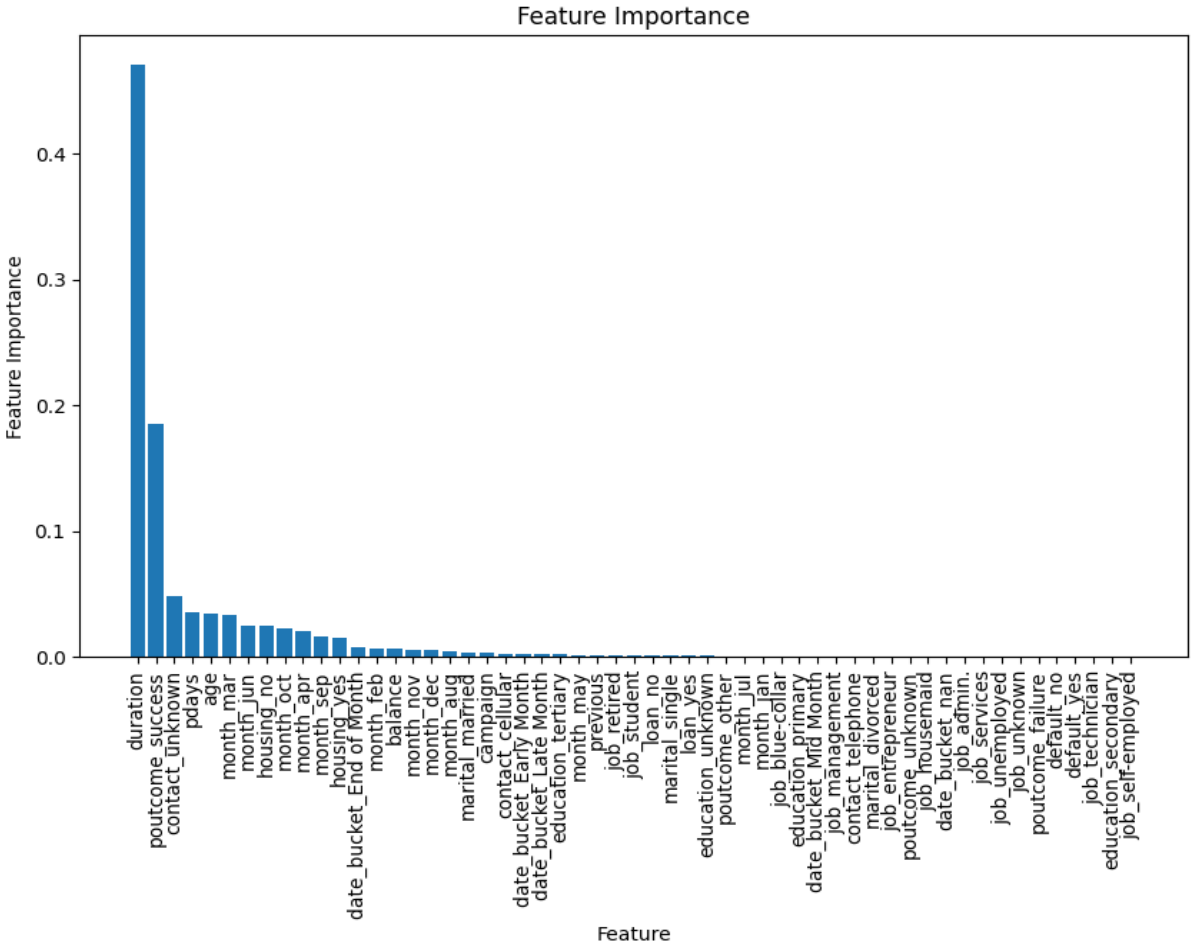

Figure 5: Feature Importance using Gradient Boosting

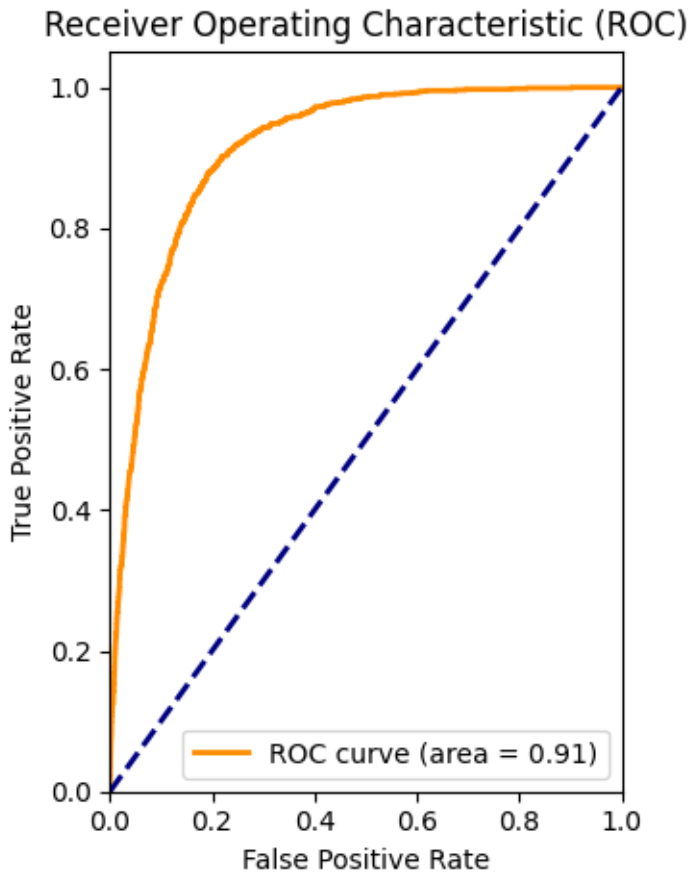

Figure 6: ROC curve for test data

Conclusion & Learning

Our data-driven journey began with a deep dive into understanding our dataset, where we uncovered valuable insights. We ensured data quality and prepared it for modeling, crafting a new feature to enhance our predictive capabilities. Leveraging advanced modeling techniques, Gradient Boosting emerged as our champion with remarkable accuracy.

This model's success was further validated through feature importance analysis and a thorough examination of its predictive performance using metrics like the Receiver Operating Characteristic (ROC) curve. Our ROC curve demonstrated the model's prowess, yielding a near-perfect area under the curve (AUC) score.

In essence, our model is well-equipped to tackle the problem at hand. Its ability to make accurate predictions and the intuitive feature engineering efforts employed make it a promising tool for decision-making. With this, we conclude this report, paving the way for informed and data-driven choices in our endeavors.

Future Enhancements

1. Feature Engineering: Continue to explore and experiment with feature engineering. New features or transformations may provide even more valuable information for your predictive models.
2. Advanced Modeling Techniques: Investigate more advanced machine learning algorithms, such as deep learning models (e.g., neural networks) or ensemble methods, to potentially improve predictive accuracy.
3. Hyperparameter Tuning: Fine-tune the hyperparameters of your chosen model to optimize its performance further. Techniques like grid search or random search can be employed for hyperparameter optimization.
4. Model Interpretability: Utilize techniques and tools for interpreting and explaining the decisions made by your model. This can enhance trust and understanding of the model's predictions.
5. Automated Machine Learning (AutoML): Consider using AutoML platforms that automate the process of model selection, feature engineering, and hyperparameter tuning, saving time and potentially improving model performance.

References

1. https://github.com/ydataai/ydata-profiling
2. https://scikit-learn.org/stable/supervised_learning.html#supervised-learning
3. https://archive.ics.uci.edu/dataset/222/bank+marketing
4. https://pandas.pydata.org/docs/user_guide/index.html
5. https://matplotlib.org/stable/plot_types/index.html