NOVA SCHOOL OF BUSINESS & ECONOMICS

# Group Project

# 2487 - Machine Learning - 2324

*Submitted to:*
Nuno André da Silva
Guilherme Brites Ramos

*April 3th, 2024*

*by:*
Carolina Marques Do Santos Aparício - 61582
Dewang Ban - 58200
Wale Doura - 57555
Erica Francalanci - 61600
Júlia Zhou Cheng – 60490

## Abstract

In this report, our objective is to leverage machine learning to tackle the pressing issues our voucher service company is confronting, namely the financial losses and complaints from our partner merchants stemming from customer no-shows. We initiated our analysis with Exploratory Data Analysis (EDA) to discern customer preferences, which informed the construction of four classification models. These models were assessed using confusion matrices and comparative analysis. Drawing on the insights gleaned from feature importance, we offer targeted recommendations for refining daily operations and enhancing the efficiency and effectiveness of our business model.

## Introduction

CouponSavvy specializes in distributing coupons for a diverse array of services, including dining, beauty treatments, and various leisure activities. By collaborating with businesses, CouponSavvy strives to offer compelling deals that enhance both sales and customer satisfaction. Despite these efforts, a substantial challenge has emerged: merchants are incurring financial losses due to customers who make reservations with coupons but fail to honor them.

As a preeminent and respected agency, CouponSavvy is pressed to address the accumulating complaints from merchants, who are facing financial setbacks when customers fail to fullfill their reserved bookings. Merchants, besieged by no-show clients, are increasingly seeking reimbursements.

The CouponSavvy advisory board is dedicated to developing predictive models to identify customers likely to not fulfill reservation commitments, minimizing revenue loss and negative reviews. We've obtained the dataset from the company covering almost one-year transactions from November 11th, 2013, to November 3rd, 2014. Our objectives include visualizing customer preferences, implementing predictive models for forecasting no-show probabilities, and comparing their effectiveness. Our strategic recommendations aim to reduce merchant churning and reimbursements by canceling high-risk transactions. By leveraging data-driven insights, we empower businesses to make informed decisions, strengthening CouponSavvy's industry leadership.

## Business Problem

The primary challenge is predicting no-shows for reservations, a crucial issue as merchants face financial losses from each unfulfilled booking. They incur costs by paying our company for each reservation, regardless of customer attendance. Accurately forecasting these no-shows is vital to protect merchants' profits and maintain their trust in our service.

## Data Quality

In tackling the challenge of financial losses due to customers not honoring bookings made with coupons, we've significantly enhanced your data quality. This enhancement involves correcting errors, ensuring consistency, removing duplicates, and enriching missing information, which is vital for accurately predicting and thus mitigating such occurrences. This initiative is crucial for upholding merchant satisfaction and improving customer experiences.

Our first step was to refine your data by removing missing values and correcting inaccuracies in prices and discounts. We established strict standards to ensure that discounts and total prices were logical and uniform across your dataset, laying the groundwork for reliable analysis.

We then focused on ensuring the uniqueness of each dataset entry by identifying and removing duplicate records, retaining only the most recent versions. This process not only ensured the uniqueness of each piece of data but also reinforced the dataset's completeness and accuracy, which are key to insightful analysis.

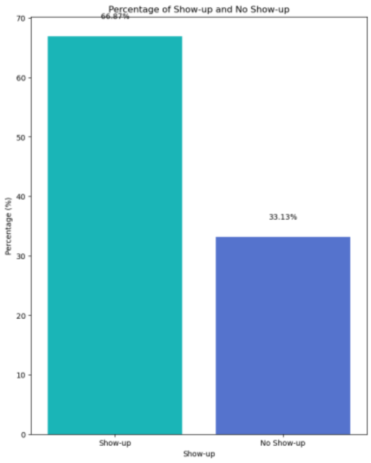
Furthermore, we validated the data against logical business operations, such as verifying that purchases within the same subcategory by a user in a brief timeframe were restricted. We also ensured that entries indicating zero attendees accurately represented instances of no-shows.

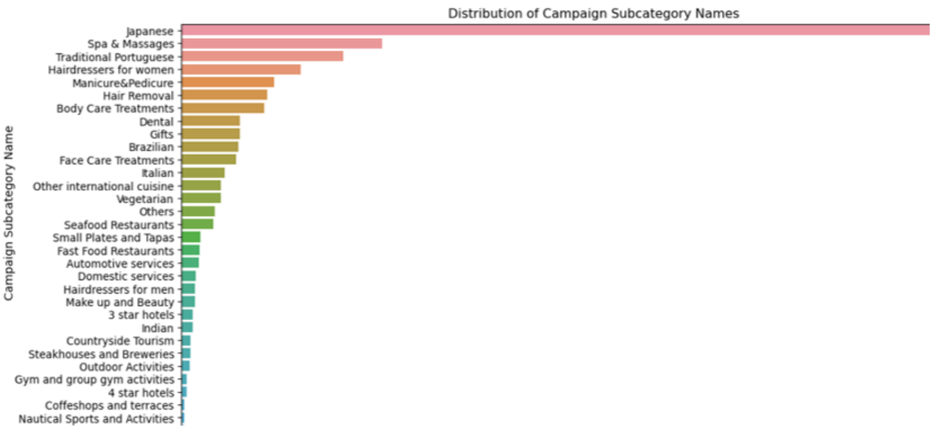
Our analysis of outliers proved crucial in understanding the distribution of transactions and identifying significant trends, preserving the integrity of the data even for non-standard entries. This in-depth analysis revealed patterns in transaction behaviors and offered insights into the distribution of statuses and values, demonstrating that a majority of transactions are valid. The status "Confirmed by lack of response" was identified as the most common, highlighting a prevalent issue of no-shows.

### Exploratory Data Analysis

In analysing the dataset from November 11, 2013, to November 3, 2014, which comprises 55,321 transactions, we've identified key insights into consumer behaviour and the company's market footprint. Our findings reveal a marked preference among consumers for web-based voucher purchases compared to mobile transactions, highlighting an opportunity for the company to further improve its web platform. This trend, coupled with Lisboa's prominence as the top county for transaction frequency, underscores the company's significant market presence.

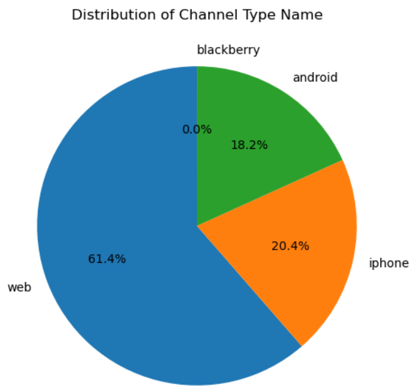
A critical concern identified in the analysis is the substantial no-show rate, with 33.13% of customers failing to honour their reservations. This issue, wherein roughly one in three customers do not show up, poses challenges for merchant satisfaction and revenue generation.

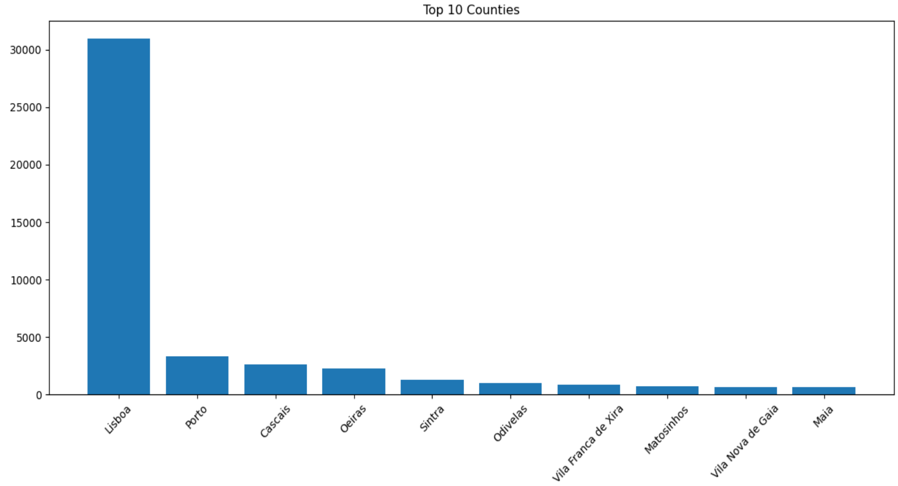



On a more positive note, the dataset also sheds light on consumer preferences, with a notable inclination towards Japanese and traditional Portuguese cuisines, along with spa and massage services, and beauty treatments such as hairdressing, and manicure & pedicure. These trends offer a window into customer interests, guiding potential marketing and service expansion strategies.

An intriguing finding is the variability in pricing for identical services within the same campaign subcategory, like the Body Care Treatments service at merchant 34th with varying charges. This suggests that merchants may offer services at different quality levels under the same category, aiming to cater to a wider range of customer preferences and budgets. This approach underlines the strategic use of differentiated pricing to appeal to diverse customer segments. The transaction data reveals most purchases are made by individuals or small groups, with a consistent spending pattern across transactions, suggesting stable customer spending behaviour. The presence of substantial discounts, with a median rate of around 50%, hints at strategies to encourage group purchases.




## Modelling

As previously mentioned, the company aims to differentiate between customers who show up for their appointments and those who do not, to align with our business objectives. The most appropriate machine learning approach for this task is a classification model. This model type is adept at predicting specific outcomes based on historical data, targeting high accuracy and reliability.

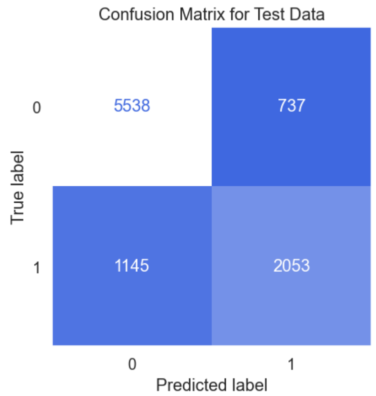
It has been chosen to train XGBoost, Decision Tree, Random Forest, and Neural Network models. XGBoost and Random Forest, both ensemble methods, are known for improving prediction stability and accuracy. Decision Trees provide straightforward, interpretable decision-making paths, whereas Neural Networks capture complex patterns with layered node structures like the human brain.

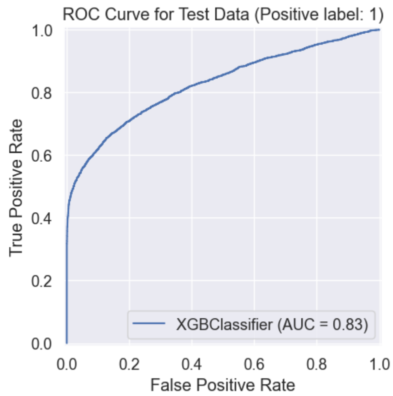
To evaluate these models' performance, we rely on metrics such as the confusion matrix, ROC curve, and metric scores including accuracy, precision, recall, F1-score, and AUC. These tools help us quantify each model's effectiveness.

Given our specific need to accurately identify no-shows due to their associated costs to merchants and potential damage to our reputation, recall—or sensitivity—is deemed the most critical metric. Recall measures the proportion of actual positive instances (no-shows) correctly identified by the model, essential for minimizing overlooked no-shows. Also, accuracy is valued for assessing the model's performance across all predictions.

Here it follows the results obtained running the different models:

### XGBoost – Model evaluation and Metrics

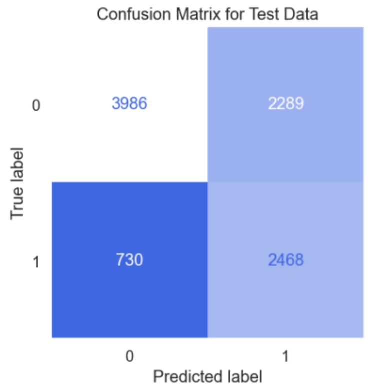


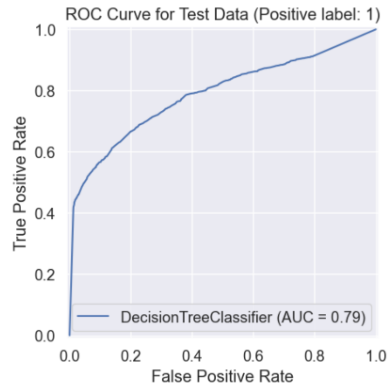




| | XGBoost |
|---|---|
| Accuracy | 80.13% |
| Precision (Class 1) | 73.58% |
| Recall (Class 1) | 64.2% |
| F1-score (Class 1) | 68.57% |
| AUC (Class 1) | 82.66% |

NOVA SCHOOL OF BUSINESS & ECONOMICS

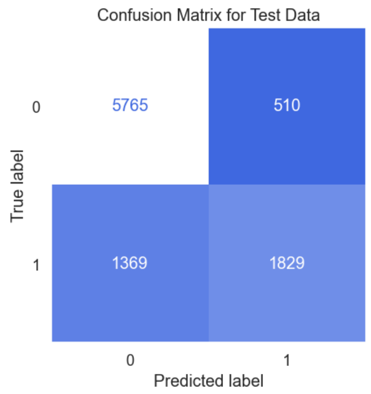
## Decision Tree – Model evaluation and Metrics

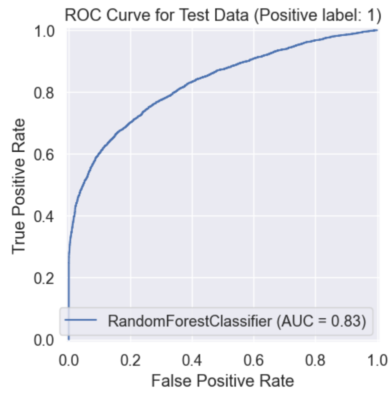



| | Decision Tree |
|---|---|
| Accuracy | 68.13% |
| Precision (Class 1) | 51.88% |
| Recall (Class 1) | 77.17% |
| F1-score (Class 1) | 62.05% |
| AUC (Class 1) | 78.94% |

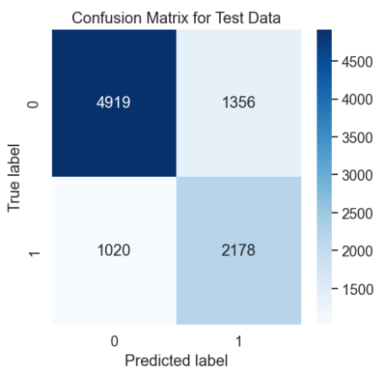
## Random Forest – Model evaluation and Metrics

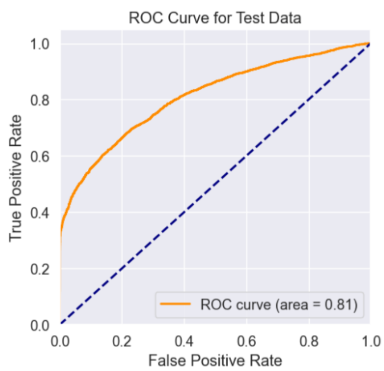



| | Random Forest |
|---|---|
| Accuracy | 80.16% |
| Precision (Class 1) | 78.2% |
| Recall (Class 1) | 57.19% |
| F1-score (Class 1) | 66.06% |
| AUC (Class 1) | 82.87% |

## Neural Network – Model evaluation and Metrics




| | Neural |
|---|---|
| AUC (Class 1) | 81.04% |
| Accuracy | 74.92% |
| F1-score (Class 1) | 64.71% |
| Precision (Class 1) | 61.63% |
| Recall (Class 1) | 68.11% |

## Model Comparison

In the comparative analysis of models to predict appointment no-shows, with an emphasis on accuracy and recall for prioritizing operational efficiency and cost reduction, the models presented nuanced but insightful results across various metrics.

XGBoost excelled with 80.13% accuracy and 64.2% recall, offering a strong balance with an AUC of 82.66%. The Decision Tree model showcased the highest recall at 77.17% but lower accuracy (68.13%) and precision (51.88%), suggesting its exceptional sensitivity despite room for improvement. Random Forest paralleled XGBoost in accuracy (80.16%) with the best precision (78.2%) but a lower recall of 57.19%, indicating precision at the cost of sensitivity. Neural Networks, set at a 0.55 threshold, demonstrated versatility with 75% accuracy and notable recall scores (78% for class 0 and 68% for class 1), proving effective in identifying both showups and no-shows.

Overall, while XGBoost and Neural Networks present balanced and promising options, the Neural Networks models excel in minimizing false positives identification. Thus, aligning with our previously stated objectives, Neural Networks is the optimal model selected for addressing our needs.

## Conclusion and Recommendations

As highlighted earlier, the marketplace model involves charging merchants for each customer's booking. This system poses a risk when customers fail to show up for their appointments, leaving merchants with costs that are not reimbursed. Consequently, merchants bear the financial burden of a service for which they receive no compensation.

In the context of our digital marketplace, maintaining the equilibrium between merchant satisfaction and customer engagement is crucial for profitability. A single failed transaction can significantly erode a merchant's trust in our platform, potentially leading to churn as they seek more reliable alternatives.

To safeguard against this, our strategy focuses on predicting customer no-shows with high accuracy. By identifying potential no-shows in advance, we can take proactive steps to minimize merchant losses and maintain their confidence in our platform. This predictive approach not only helps in retaining our existing merchant base but also positions us as a trustworthy and supportive intermediary in the digital marketplace, distinguishing us from competitors and attracting new merchants seeking a reliable booking platform.

Building on this rationale, we have prioritized a predictive model specifically designed to accurately identify customers likely to not show up for their appointments. This prioritization stems from the understanding that false negatives—incorrectly assuming a customer will attend when they do not—carry not only a direct monetary cost for our merchants but also potentially erode our company's reputation.

Consequently, we have favored a model that exhibits a high recall score. This means the model is tuned to maximize the identification of true no-shows, thus minimizing the instances of false negatives. By focusing on high recall, we aim to reduce the financial impact on our merchants and safeguard our platform's standing as a reliable and merchant-friendly digital marketplace. This strategic choice underscores our commitment to supporting our merchants' success and upholding the trust they place in our platform.

Therefore, we advise deploying the Neural Network model to effectively reduce the no-show rate, thereby diminishing merchant churn, and curtailing adverse social media feedback about our company. This model promises a robust 75% accuracy and a 68% recall score, which are commendable metrics, especially without customer demographic data. It leverages purchase dates, categories, channels, and price charged for its predictions.

A detailed examination indicates that the purchase date, notably the month, is a key predictor of no-show incidents. Therefore, we recommend increasing oversight on transactions in months identified as having a higher risk of no-shows. Furthermore, the fees charged to merchants and the customer-chosen subcategories are significant factors affecting no-show rates. We suggest enacting specific measures, such as tailored follow-ups in categories with lower fees or imposing deposit requirements on sought-after items, like restaurant deals, to bolster compliance with reservations and enhance merchant contentment.

We propose establishing a data-informed compensation model for merchants encountering revenue loss due to no-shows, which will aid in maintaining and strengthening our business relationships. Our commitment to refining the predictive model with current data will boost its accuracy and dependability. The execution of these strategies is crucial in reinforcing CouponSavvy's position as a leader in the market.