UNC Universidad Nacional de Córdoba

Diplomatura en Ciencia de Datos e

Inteligencia Artificial y Sus

Aplicaciones en Economía y

Negocios

Data Science & Artificial Intelligence Diploma
Data Applications Module

Predictive Model for Marketing Strategies in Tourism E-commerce

Tutor: Mgter. Giuliano Fassano

Team members:
Albornos, Juan Martin
Fogliatto, Robertino
Oviedo, Rocío Nazareth
Zurita, Emilia

1. Introduction

This report presents the development of the assignments for the “Applied Integration Module,” whose purpose is to practically apply the knowledge acquired throughout the Data Science and Artificial Intelligence Diploma and its applications in Economics and Business.

The project details the complete development of a predictive model to estimate users’ propensity to purchase tickets. The process includes all stages of the data science workflow: from dataset exploration and cleaning, to the implementation and comparison of different classification models. The evaluated models include Logistic Regression, Decision Trees, Random Forest, and XGBoost.

User behavior analysis in digital platforms is a key field within data science applied to marketing. Companies seek to understand which variables influence conversion and what behavioral patterns are associated with purchasing decisions. In this context, the objective of the project is to build a classification model capable of predicting whether a user will purchase a ticket based on a dataset provided by the course.

2. Dataset Description

The original dataset was obtained from a Google Drive source and contains information about users of a travel platform. It includes numerical and categorical variables related to user interaction with the website—such as the average number of views, likes, and check-ins—along with demographic and behavioral characteristics.
The dataset contains 11,760 records and 17 columns, describing both user activity and contextual variables.

3. Data Processing and Cleaning

The first step was identifying and correcting inconsistencies in the data. An initial analysis using df.info() revealed the presence of missing values across multiple columns and data types that did not match the nature of the underlying information. An exploration of unique values exposed various typographical errors.

Column-wise transformations were applied to correct these issues. Examples include:

- Replacing the symbol “*” with null values in yearly_avg_Outstation_checkins and converting the column to Int64.
- Replacing “Three” with the numeric value 3 in member_in_family and converting the column to Int64.
- Correcting the value “Yeso” to “Yes” in following_company_page.

In the column Preferred_device, the same device category appeared under different names. We grouped all mobile devices under the label **“Mobile”**, since “ios and android” was the broadest category and did not allow distinguishing OS-level differences.

To detect outliers, boxplots were generated for numerical variables. Several variables showed numerous extreme values. To treat these, we used the 5th and 95th percentiles and removed values exceeding the 95th percentile plus 1.5 times the IQR. No lower threshold was applied, since values below that range could naturally occur.

For missing values, we first quantified nulls per column and then applied imputation strategies based on the nature of each variable: mean for normally distributed variables, median for skewed ones, and row deletion when the proportion of nulls was minimal.

After completing the cleaning process, we obtained a dataset free of typographical errors, missing values, and outliers.

4. Exploratory Data Analysis (EDA)

The exploratory analysis provided a comprehensive view of user behavior. We generated histograms for continuous variables and bar plots for categorical variables. Key observations include:

- Most users had not purchased tickets, confirming class imbalance.
- Annual average views on the travel page showed a symmetric, nearly normal distribution, concentrated between 200 and 350.
- Tablets and mobile devices were the most common; laptops were scarce, and “Other” was nearly residual.
- Likes were highly dispersed, with a peak near 30,000.
- Check-in activity was low, concentrated between 0 and 2 per user.
- Small families (1–4 members) predominated.
- Preferred destinations clustered around Beach, Financial, Historical Site, and Medical.
- Comments were concentrated between 50 and 95 per year.
- Total_likes_on_outstation_checkin_received showed strong right skew.
- Most users had recent check-ins (0–3 weeks).
- More than twice as many users did not follow the company page.
- Comment activity within the company page was low overall.
- “No” responses in Working_flag were nearly five times more common than “Yes.”
- Travelling_network_rating was concentrated between 3 and 4.
- Adult_flag values were mostly 0 or 1.
- Daily time spent on the page ranged mainly from 8 to 20 minutes, with few users exceeding 30–40 minutes.

Segmenting histograms by purchasers vs. non-purchasers revealed that buyers tend to have:

- Higher average page views
- More annual check-ins
- Slightly longer daily browsing times (30–60 min)
- Shorter intervals since last check-in

The correlation matrix showed positive associations between time spent on the page and likes received, consistent with higher activity. Recent check-ins also correlated moderately with views and time spent.

Contingency tables by Taken_product revealed:

1. Users who follow the company page show higher purchase propensity
2. Mobile and laptop users have lower purchase rates
3. Destination category “OTT” shows the highest buyer proportion, followed by “Tour and Travel” and “Hill Stations”
4. “Other,” “Medical,” and “Game” have the lowest proportions

5. Principal Component Analysis (PCA)

PCA helped identify global patterns and reduce dimensionality; however, class separation was limited. Explaining 80% of the variance required eight components, suggesting:

- The relationship with the target variable is likely non-linear
- Quantitative variables are not strongly correlated
- Complex modeling techniques are necessary to capture meaningful patterns

Thus, more sophisticated classification models were required.

6. Predictive Models

During the modeling stage, four classification models were built and evaluated: Logistic Regression, Decision Tree, Random Forest, and XGBoost. Each model aimed to predict the target variable *Taken_product* (converted to a binary format: No = 0, Yes = 1) using the full set of predictor variables, excluding *UserID*.

In all cases, preprocessing was carried out using a *ColumnTransformer*, which included standardizing numerical variables with *StandardScaler* and encoding categorical variables using *OneHotEncoder*. Given the class imbalance, weighting strategies were applied—*class_weight="balanced"* in linear and tree-based models, and *scale_pos_weight* in XGBoost—to assign a stronger penalty to misclassifications of the minority class. Model validation was performed through five-fold cross-validation, and accuracy, precision, recall, F1-score, and ROC-AUC were calculated on both training and test sets to ensure the robustness of the results. Additionally, *GridSearchCV* was used to determine the optimal hyperparameters for each model.

Logistic Regression, used as the baseline model, achieved an F1-score of 0.464, a recall of 0.759, and a ROC-AUC of 0.804. Coefficient analysis revealed that the most influential variable was whether the user follows the company page, which increases purchase likelihood by a factor of 1.17. Despite its simplicity and interpretability, the model produced a high number of false positives and was unable to capture the complexity of the relationships between variables. Thus, although useful as an initial benchmark, its linear nature makes it insufficient for accurate and reliable predictions. This justified the need for more complex models, such as decision trees or ensemble learning methods like Random Forest and Gradient Boosting.

The **Decision Tree** model, after hyperparameter tuning with GridSearchCV, achieved much higher metrics on the training set (F1-score of 0.92 and ROC-AUC of 0.953). Although this suggested potential overfitting, performance on the test set was even higher. An analysis of F1-score curves for training and validation across different tree depths showed slight overfitting beginning around *max_depth = 12*, which led us to set this as the final depth.

Data leakage was also ruled out by confirming that no feature was directly derived from the target variable. Therefore, the model functioned correctly and classified users effectively.

Using the refined model with *max_depth = 12*, we identified the threshold that maximized the F1-score. The optimal threshold was 0.813, meaning users were classified as buyers only when their predicted probability exceeded that value. This made the model more conservative and improved the balance between precision and recall.

As a result, when the model predicted that a customer would buy (class 1), it was correct 89% of the time and successfully detected 89% of actual buyers. From a marketing perspective, this is highly advantageous, as it minimizes unnecessary promotions to users who would not purchase while ensuring potential buyers are effectively identified.

The Precision–Recall curve was also high, indicating that many true buyers were ranked near the top of the model's probability distribution. The ROC curve was excellent, reflecting strong class separation. The most important variables for class discrimination were *total_likes_on_outofstation_checkin_received*, *total_likes_on_outstation_checkin_given*, and *Yearly_avg_view_on_travel_page*.

The **Random Forest** model improved overall stability and delivered stronger performance on the test set, achieving an F1-score of 0.956 and a ROC-AUC of 0.999. It showed a slight imbalance between precision (0.991) and recall (0.92), with the former being extremely high and the latter slightly lower.

The threshold that maximized the F1-score on the training set was 0.385, indicating that the model tended to assign moderately low probabilities to the positive class. Using the standard 0.5 threshold would have caused many positives to be missed. Lowering the threshold reduced precision to 0.945 (still very high) and increased recall to 0.965, achieving a strong balance between the two. The three most important predictors were the same as in the Decision Tree.

Overall, recall improved relative to the tree model, now reaching 0.965, and precision and F1-score also increased, yielding an excellent trade-off between identifying buyers and avoiding unnecessary contacts.

Both the ROC and Precision–Recall curves confirmed the model’s outstanding ability to distinguish between classes.

Finally, the **XGBoost** model outperformed all previous models by combining the strength of decision trees with advanced regularization techniques. Its ability to build an ensemble of trees and automatically adjust penalties for imbalanced classes resulted in a more robust and generalizable model. It performed exceptionally well, showed no signs of overfitting, and improved upon Random Forest’s metrics, achieving recall of 0.976, precision of 0.989, and an F1-score of 0.982. With only 4 false positives and 9 false negatives, class boundaries were learned with remarkable accuracy.

The optimal threshold maximizing the F1-score was 0.595, slightly higher than the standard 0.5 to make positive classifications more conservative. However, adjusting this threshold had almost no impact on the model’s performance, as precision increased minimally (to 0.992) and the remaining metrics remained practically unchanged. The model’s test performance (F1 ≈ 0.984) was slightly higher than the training performance, suggesting no overfitting.

This indicates that XGBoost was already extremely well calibrated, and the default threshold of 0.5 was near the optimal trade-off between precision and recall.

Both ROC and PR curves achieved a score of 1, confirming the model’s exceptional discriminative power. The most important individual predictors were whether the user follows the company page, *Adult_flag*, and “Other” as the preferred destination category, while the most influential aggregated predictors were *Adult_flag*, *total_likes_on_outofstation_checkin_received*, and *travelling_network_rating*.

7. Model Evaluation and Comparison

The model comparison shows a clear progression in performance. Logistic Regression provides a solid interpretative baseline, whereas ensemble models such as Random Forest and XGBoost demonstrate a significant improvement in predictive capability. The incorporation of cross-validation techniques and hyperparameter tuning helped mitigate the risk of overfitting and ensured better model generalization.

Regarding interpretability, the most influential variables in the final models were those related to digital interaction and affinity with the company—specifically, following the page, preferred destination category, and daily usage time. These findings are consistent with expected behavior in digital marketing contexts, where active engagement and brand connection are strong indicators of higher purchase likelihood.

Therefore, XGBoost is selected as the model for predicting purchase propensity, as it offers the best overall performance, excellent generalization capacity, and interpretations that align with the behavioral patterns observed in the data.

8. Final Analysis

We analyzed the false negatives, meaning users who were not identified by the model as buyers but ultimately completed a purchase. In this case, the model tends to underestimate them. We observed that all of these users accessed the platform from mobile devices, and that 6 out of the 9 shared the same value in the *likes_on_outstation_checkin_given* variable. Additionally, the *working_flag* field was "No" for all of them.

We then examined the false positives, meaning users predicted as buyers who ultimately did not complete a purchase. This group is particularly important, as they represent users for whom targeted marketing actions may help reverse their decision. All of these users also accessed the platform from mobile devices, and 4 out of the 5 presented the same number of *likes_on_outstation_checkin_given* (matching the values observed in the false negatives), while *working_flag* likewise took the value "No."

Based on these observations, there is no evidence of a specific group of users with particular characteristics that the model systematically misclassifies.

We then divided the dataset into deciles to analyze the predicted purchase probability (y_proba) within each group.

The results show that:

- The lowest 10% of clients (decile 1) have values below 0.0007, which are essentially zero.
- The highest 10% (decile 10) exhibit probabilities above 0.98, indicating an extremely high likelihood of purchasing.
- The overall distribution is highly concentrated: most clients display low purchase probability, while a small group shows very high probability.

This suggests that the model has excellent discriminative power, clearly separating buyers from non-buyers—a finding consistent with the ROC-AUC and PR-AUC metrics, both equal to 1.

Based on this information, we conclude that

- Purchase probability is practically null for deciles 1 through 5, meaning that any marketing effort toward these users would be inefficient.
- Probability increases slightly in deciles 6 through 8, but remains low, so resource investment in these segments is also not recommended.
- It is likewise unnecessary to allocate efforts to decile 10, since purchase likelihood is already nearly guaranteed.

Thus, the main value provided by the model lies in its ability to accurately identify the 80% of users who will not purchase and the 10% who almost certainly will. The next step is to examine decile 9 in detail, as it contains users with intermediate purchase probability. Within this group, sub-deciles 4 and 5 appear to be the most suitable targets for marketing actions aimed at stimulating conversion.

9. Application of Model Results to Marketing Strategies

Based on the results obtained from the predictive model, we propose designing a marketing campaign primarily aimed at undecided users, with the goal of maximizing conversions within this segment of potential customers. The analysis showed that purchase propensity is mainly associated with the level of digital interaction, reflected in whether the user follows the company page, the number of likes and check-ins, average browsing time, and the preferred location type recorded on the platform.

In light of these findings, we recommend implementing segmentation and personalized content strategies targeted at the most active users, strengthening their connection with the brand through remarketing campaigns, notifications, or exclusive promotions aligned with their behavioral patterns. It is also advisable to use the predictive model on an ongoing basis to continuously update user segments and monitor the impact of the implemented actions through conversion and retention metrics. In this way, the integration of data analysis and personalized marketing will make it possible to optimize available resources and enhance the acquisition of new customers.