machine learning report

Supervised learning project

Howest Applied Science (AI)

Denis Topallaj - Jean-Baptiste Maene – Lander Pauwels Mallengier

Contents

[1. Background and Introduction 2](#_Toc129635244)

[2. Data collection and Preparation 2](#_Toc129635245)

[3. Methodology 2](#_Toc129635246)

[4. Results 4](#_Toc129635247)

[5. Future Work 4](#_Toc129635248)

# Background and Introduction

As a data scientist, it is our responsibility to analyze data and draw insights that can help businesses and organizations make informed decisions. In this project, we were tasked with screening Smurfs clients who are applying for an exclusive hotel in the Bahamas.

Smurfs are notorious for their rowdiness, unbridled spending, and tendency to cause extensive damages to hotel infrastructure and staff. These costs are typically difficult to recover once the guest has left the premises. Therefore, the objective of this project is to create a screening mechanism that can identify high-risk Smurf clients, preventing them from being admitted to the hotel. By doing so, we aim to minimize potential losses and maintain the hotel's reputation for excellence.

In the following sections, we will discuss the methodology we used to develop this screening mechanism, the results we obtained, and recommendations for the hotel management based on our findings.

# Data collection and Preparation

Creating an AI that can determine the best and worst Smurf, needs data. We collected data on Smurf clients who had stayed in the hotel in the past, including information about their demographics, past behavior, and any damages they caused.

This given data was then transferred to two files named “train.csv” and “score.csv”. These files were not ready to be used on the models. They contained empty rows, duplicates, outliers, or sometimes just wrong information. These rows have been removed or modified from the dataset. Categorical features have been split up into two columns. Features such as ‘spa’ have been split up into “yes\_spa” and “no\_spa”. Columns that contained true or false values were transformed into 1’s and 0’s. These operations were conducted to ensure that the models have enough accurate information.

# Methodology

To develop a screening mechanism for high-risk Smurf clients, we used a machine learning approach that involved building a predictive model based on historical data.

The goal of this project is to determine the profit of the last visit, if the Smurfs have caused damage and the cost of the damages if they did so. We used the cleaned datasets with several machine learning algorithms including polynomial regression, decision tree, K-Nearest Neighbor, Gradient Boosting Regression and many more. For the prediction of the profit for the last visit we used Random Forest Regressor.

A picture containing chart

Description automatically generated

Figure 1: Feature influence outcome profit

We can see on this graph (Feature 1) that the most influential feature is the profit\_am (total profit across visits (excluding last visit)). This accompanied by the number of nights booked are a good indicator of Smurfs that are big spenders.

Smurfs also needed to be classified inside two groups: “caused damage” or “didn’t cause damage”. For this reason, we had to use classification models such as logistic regression, random forest, support vector machines and many more.

Chart

Description automatically generated

Figure 2: Plot models classification

The picture (Figure 2) shows the best models that were created with our dataset. Logistic regression and Polynomial regression were the best ones. The logistic regression model was used on our test dataset to predict which Smurfs would cause damage to the property.

Lastly, we also need to predict the cost of the damages that Smurfs would cause. The same principles were used such as the prediction of the outcome\_profit. The next picture (Figure 3) shows the influence of features related to the damage amount per Smurf. We can clearly observe that the more a Smurf uses the bar, the more influence this has on the damage amount. As stated in the section “Background and Introduction”, Smurfs are known for their rowdiness. Having a drink or two could cause more damage to the hotel.

A picture containing chart

Description automatically generated

Figure 3: Feature influence damage amount

With a lot of tuning, we could not create a fitting model for this feature. The best testing score available was 6%. This could be due to a lot of reasons such as insufficient or poor-quality data, imbalanced data, incorrectly tuned hyperparameters, overfitting and much more.

In summary, our methodology involved using cleaned data that was collected and preprocessed, selecting a suitable machine learning algorithm, building, and tuning the model, and testing this on unseen data. These predictions were then added to the “score\_cleaned” file.

# Results

In this section, we present the results of our machine learning model for predicting the best Smurfs. We filtered the data based on the best possible attributes of a good Smurf. This includes being a big spender, having a high neighborhood income, the amount of nights booked, the profit generated and much more.

This data was also filtered and sorted on a negative way for example the amount of damage caused, the amount of times a Smurf went to the bar, the negative score from the staff etc.

The ideal Smurf is a 45-year-old (male or female) who has around 2 adults in their family and is mostly childless. They booked around 17 nights and requests a drink at the bar an average of three times. This Smurf likes sports, shopping, using credit cards and is on retirement. On their last visit they bring in a profit of €1919 with a little bit of damage that results to around €24.49.