NOCTUA LIBRARY PITCH

INTRODUCTION

What does the new Chips Ahoy marketing director do her first day on the job? Enable cookies

Hello, we are Dux, and we are a marketing agency specialized in helping entrepreneurs to start their business, adjusting to their budget with the help of new technologies such as machine learning instead of the typical market studies that are the focus groups, interviews, etc. Our motto is: *"We will find a way"*.

To explain a little better who we are, we will tell you about one of our most recent clients, Noctua, the owners told us that according to an investigation carried out by INEGI around 72% of adults in Mexico read books, so they believe that setting up a bookstore would be a great investment idea.

Unfortunately, they do not have a very large budget and can only allocate 120 thousand dollars to the purchase of the books they are going to sell:

1) They have already allocated 100 thousand dollars to buy existing books that are already popular, such as: *"Harry Potter"*, *"Don Quixote de la Mancha"*, *"Les Miserables"*, etc.
2) The remaining 20 thousand dollars will be allocated to the purchase of new books that do not have a history yet.

But how do you know which books to buy? That's where we intervene.

OVERVIEW OF THE PROJECT

We believe a market Study can be as good as the data used in its preparation, this is why we reviewed several online basis, from libraries and publishers data sets to social media sets.

After detailed insopection, we selected Good Reads' data set as the main reference mainly due to the different book's characteristics included in it.

NOCTUA LIBRARY PITCH

Note Good Reads' users are able to rate any book from the catalog based on their experience providing a variable that can be used as a proxy of a book's success.

The data preparation phasis included:

1) Basis merging,
2) Data cleaning, and
3) Categorical variables treatment to make them useful for our analysis.

During data preparation phasis, we were concern about potential bias in the rating variable if, for any reason, customers tend to rate only books they like or dislike, thus, we created a visual analysis to show that rating seems to be normally distributed.

Finally, during testing phasis, we reviewed 4 different methodologies: Random Forest, Logistic Regression, Neuronal Network with Single and Multiple layers and decided to use Logistic Regression due to its performance.

DATA ANALYSIS

Dependent Variable:

We have a normal independent variable based on the historic books' ranking.

Continuous Independent variables:

We evaluated two continuous variables as independent variables: Price and number of pages. Note both variables are naturally a little bit skewed to the right.

Categorical Independent Variables:

We included 5 categorical variables:

- Author.
- Publisher.
- Category.
- Language.
- Part of a series.

Note all of these variables are very pulverized as can be seen in the graph showing publisher's dispersion, thus, we created new variables by grouping them in 2 or 3 new categorical variables.

NOCTUA LIBRARY PITCH

LOGISTIC REGRESSION

Finally, we implemented a logistic regression model that will output a binary result which, on this case, is the dependent variable (rating) in function of the independent variables described before (both continuous and categoricals).

To create the binary dependent variable, we defined a threshold parameter to tell us if the book will be successful or not:

1) Ranking measure being 4.1 and above was considered as a 1 (successful).
2) Ranking 4.1 and bellow was considered as a 0 (unsuccessful).

After reviewing the results in our confusion matrix we realized that the scores of the precision, F1 and recall are higher on identifying the 0s instead of 1s, thus, we decides to base our evaluation in the 0s meaning forecasting books expected to be unsuccessful.

Logistic Regression shown a 70% confidence, this means, if there are 100 rejected books 70% of them are expected to be fair no buy recommendations to the client.

The risk related to wrongly reject a new book to offer is that the library misses to acquire a title that may become successful, although, in this case our client will just need to buy those new hot selling books during the weeks after launch.

In conclusion, with this project we chose to provide our client a supervised machine learning algorithm forecasting which books not to buy, that way we are able to make a solid recommendation. Rather than recommending which books to buy which would lead to a large percentage of error as shown in the results of the confusion matrix.

NOCTUA LIBRARY PITCH

FINAL REMARKS

Based on the analysis presented before, Noctua's request to provide a strategy to face investment for new books offered to the library would be as follows:

1) When receiving new books to be included in the catalog, let's first use our model to identify the ones that should not be purchased due to its low chance to become success and create low value stock.
2) Let's keep reviewing sales for new books to identify any potential wrong rejection and purchase it to be included the library stock.
3) We must keep training the model when new records are available to ensure it evolves with the industry.

Finally, we can keep supporting our client, as an example, when Noctua has defined the final location for the library, we can go one step ahead and analyze potential customers in the zone to better understand their preferences and adjust the catalog accordingly.

Don't hesitate that in order to support you: *we will find a way*.