Capstone Project Proposal

Domain background

Credit risk's management is one of the most important fields in the Bank's functions, because it helps to decide to whom it's going to lend money and to whom it's not, and in case of accepting credit requests being able to determine how much money it's going to lend. In this way, Banks are always building analytics models based on characteristics of people such as historical credit defaults, demographic features, transactions features or Credit bureaus features to determine people's probability of default and people's income. These factors, according with the literature, are essential elements to effectively manage the credit risks.

In such a way, getting new variables to be more precise in the prediction of credit risk's factors is not a minor problem and can generate economic benefits. As follows, certain efforts have been made in order of obtaining and analyzing public data available, like Twitter posts (after getting the client consent to use this information in the credit analysis).

According with certain academic documents like "Predicting individual-level income from Facebook Profiles" by Matz et al, the way that people write can predict their level of income. In this way twitter posts, or any written information about the clients can be transformed in a variable which predicts the person's probability to have a high income, this new variable can be used with the already known characteristics to get a final income for each client.

Problem statement

Certain Bank in Colombia is trying to strength its Credit Risk Management models using client's tweets, after getting the consent of the consumers for use that information. The Bank will predict the level of income (lower, higher) income based on client's tweets.

Datasets and inputs

People's tweets for three or four weeks using the twitter REST API in certain zones from the five mayor cities in Colombia (Bogotá, Cali, Medellín, Barranquilla and Cartagena), labeling each tweet as "high income" or "low income" depending which zone the tweet was made. The zones were chosen according with income level polls for zones in the cities, extracting only information in the highest and in the lowest income zones.

Solution statement

Based on the information and the label for each tweet, it will be constructed a classification model that predicts if certain tweet is from a “high income” zone or from a “low income” zone.

The information will first be preprocessed in order to obtain “bag of words” variables, spelling variables and sentiment variables for each tweet, choosing the lesser possible variables in the final solution of the problem.

The model will be trained, proving different classification algorithms, comparing its metrics and choosing the best one.

Benchmark model

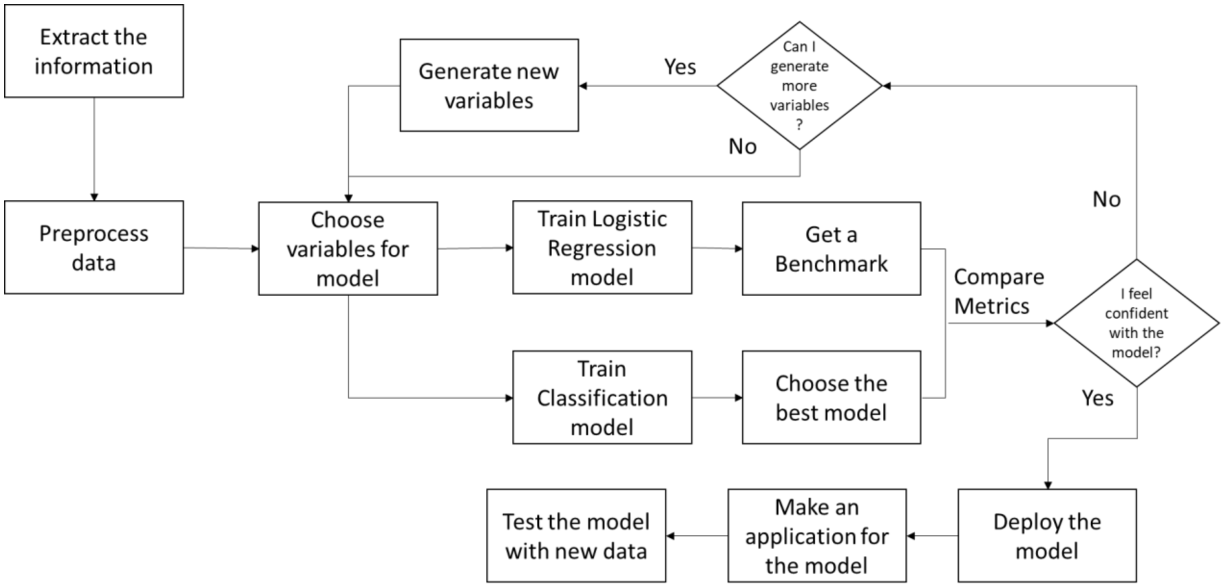
The Benchmark model would be the model obtained using a logistic classification.

Evaluation metrics

The evaluation metric would be the AUC and based on the AUC will be chosen the final model.

For the Bank perspective is equally important to have false positives because these clients can be potential defaulters in the future, or false negatives because with these clients the bank is losing a clear opportunity, so the final metric for the model will be F1.

Project design


Extract the information
Preprocess data
Choose variables for model
Generate new variables
Yes
Can I generate more variables ?
No
Train Logistic Regression model
Get a Benchmark
Train Classification model
Choose the best model
Compare Metrics
No
I feel confident with the model?
Yes
Deploy the model
Make an application for the model
Test the model with new data

References

Matz SC, Menges JI, Stillwell DJ, Schwartz HA (2019) Predicting individual-level income from Facebook profiles. PLoS ONE 14(3): e0214369. https://doi.org/10.1371/journal.pone.0214369