

Artificial Intelligence Design and Implementation

**Marketing classification Problem**

**AI Algorithms 1**

**(**AIDI 1002**)**

**Statement of Work (SOW)**

**(**01 November 2020**)**

Course facilitator: Marcos Bittencourt

Sanket Patel (100800990)

**Link:** [**https://td-subscription-prediction-api.herokuapp.com/**](https://td-subscription-prediction-api.herokuapp.com/)

**Executive Summary:**

It is extremely useful to have a system that can easily recognize patterns and behaviors, when it comes to targeted marketing, and then react accordingly. This ability to "read between the lines” is what makes machine learning so important today, and in the years to come, in marketing.

In the past, on nothing more than guesswork, many advertisers launched advertising campaigns. A lot of money was spent on advertising or promotional campaigns that did not align with their target audiences without understanding their audience.

This marketing waste is reduced by machine learning.

In the modern age, taking a scattershot approach to marketing is not only futile, but mere folly. Machine learning takes the guesswork out of the equation, enabling marketers to reach their audience with content and product offerings that are the best chance of interaction, and finally, conversions.

**Rationale Statement:**

Instead of using a scattershot approach, a targeted marketing solution not only reduces marketing costs, but also optimizes resources. The purpose of this project is to develop a solution for a targeted marketing campaign for potential customers of a banking product (term deposit) in order to optimize marketing costs.

**Problem statement:**

To classify if a customer will subscribe to term deposit or not, compare different models (Logistic Regression, Decision tree, Random Forest, SVM, KNN, ensemble models(Bagging/boosting) and ANN etc.) based on their performance, identify the best solution and deploy it in practice.

**Data requirements:**

* Data must be the banking data containing client information, previous campaign information and socio-economic data.
* Each row must be labeled into either of the two categories (“Subscribed” or “Didn’t subscribe”)
* Dataset should be unbiased (If biased, used oversampling or under sampling techniques).
* Dataset should be correctly labeled.
* For better computation, data should be sampled.
* Data should be split in train, test and validation sets.
* Sample data should represent overall population
* Dataset should adequately represent features differentiating both classes (ex.).
* No Duplicate data should be present to avoid overfitting.
* There should be no missing values.
* Data should be randomized and in no particular order.

**Assumptions**: The data is unbiased, randomized, with no exceptions, consistent and correctly labelled.

**Data:**

Data Source: <https://archive.ics.uci.edu/ml/datasets/Bank+Marketing>

This example uses data related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to assess if the product (bank term deposit) would be subscribed ('yes') or not ('no'). The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to assess if the product (bank term deposit) would be subscribed ('yes') or not ('no').

Data available:

* bank-additional.csv (Sampled data)
* bank-additional-full.csv (Full dataset, this dataset may take more time to run some of the machine learning models.)

Data Description:

1. Bank client data:

* age (numeric)
* job: type of job (categorical: 'admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown')
* marital: marital status (categorical: 'divorced', 'married', 'single', 'unknown'; note: 'divorced' means divorced or widowed)
* education (categorical: 'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown')
* default: has credit in default? (categorical: 'no', 'yes', 'unknown')
* housing: has housing loan? (categorical: 'no', 'yes', 'unknown')
* loan: has personal loan? (categorical: 'no', 'yes', 'unknown')

1. Related with the last contact of the current campaign:

* contact: contact communication type (categorical: 'cellular', 'telephone')
* month: last contact month of year (categorical: 'jan', 'feb', 'mar', ., 'nov', 'dec')
* day\_of\_week: last contact day of the week (categorical: 'mon', 'tue', 'wed', 'thu', 'fri')
* duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

1. Past Campaign Information:

* campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
* pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
* previous: number of contacts performed before this campaign and for this client (numeric)
* poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')

1. Social and economic context attributes:

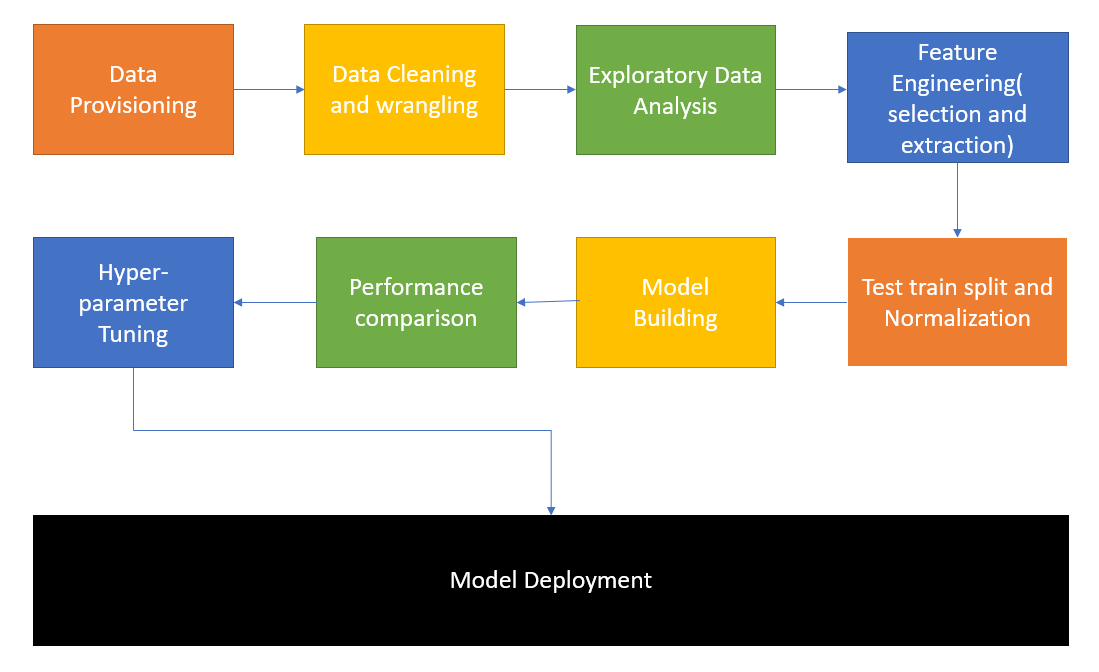
* emp.var.rate: employment variation rate - quarterly indicator (numeric)
* cons.price.idx: consumer price index - monthly indicator (numeric)
* cons.conf.idx: consumer confidence index - monthly indicator (numeric)
* euribor3m: euribor 3 month rate - daily indicator (numeric)
* nr.employed: number of employees - quarterly indicator (numeric)

5. Output variable (desired target):

* y - has the client subscribed a term deposit? (binary: "yes","no")

Missing Attribute Values: There are several missing values in some categorical attributes, all coded with the "unknown" label. These missing values can be treated as a possible class label or using deletion or imputation techniques.

**Modelling Approach:**



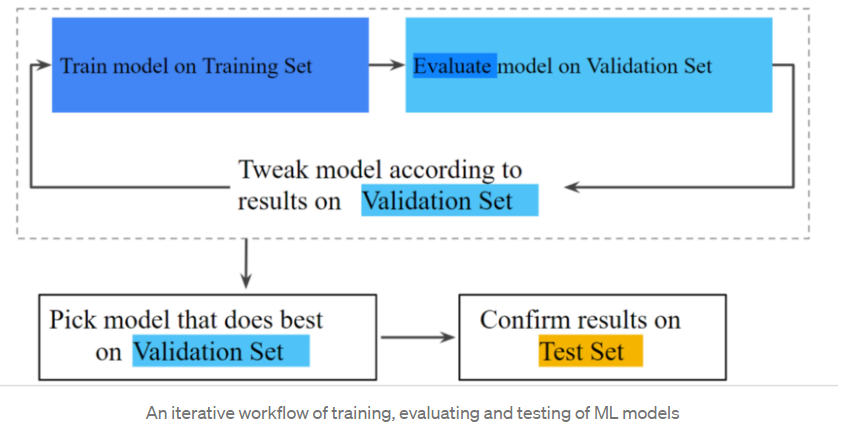
**Metrics:** Precision, recall, F1-score, Cross validation score, ROC, AUC. Accuracy.

**Tools:** Jupyter Notebook/ Google Colab

**Language:** Python3

**Libraries:** Scikitlearn, statsmodels, Tensorflow/Keras, Numpy, Pandas, matplotlib, seaborn, sweetviz/pandas profiling.

**Test procedure and model acceptance:**



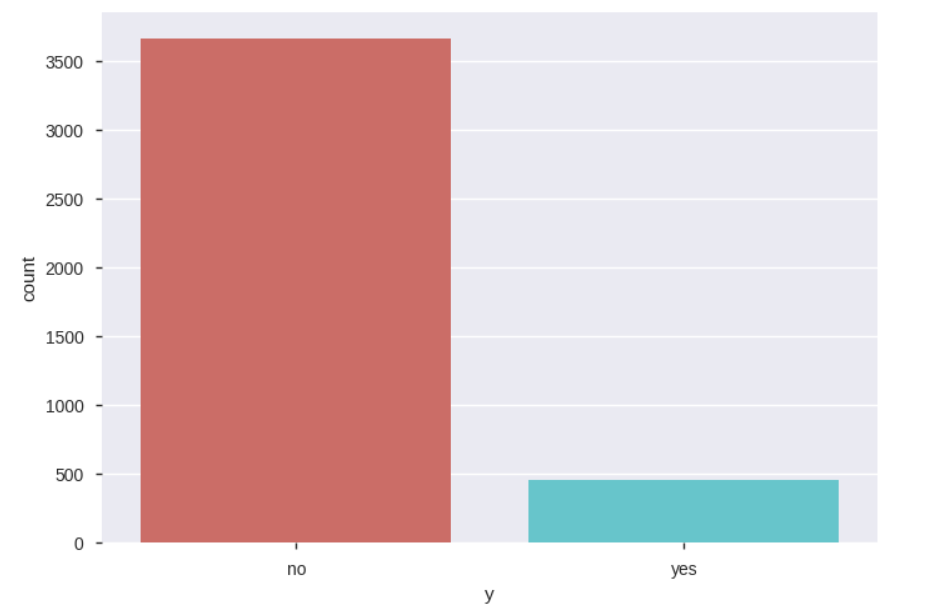
Source: <https://medium.com/analytics-vidhya/testers-guide-for-testing-machine-learning-models-e7e5cea81264>.

For evaluation following techniques/metrics will be used:

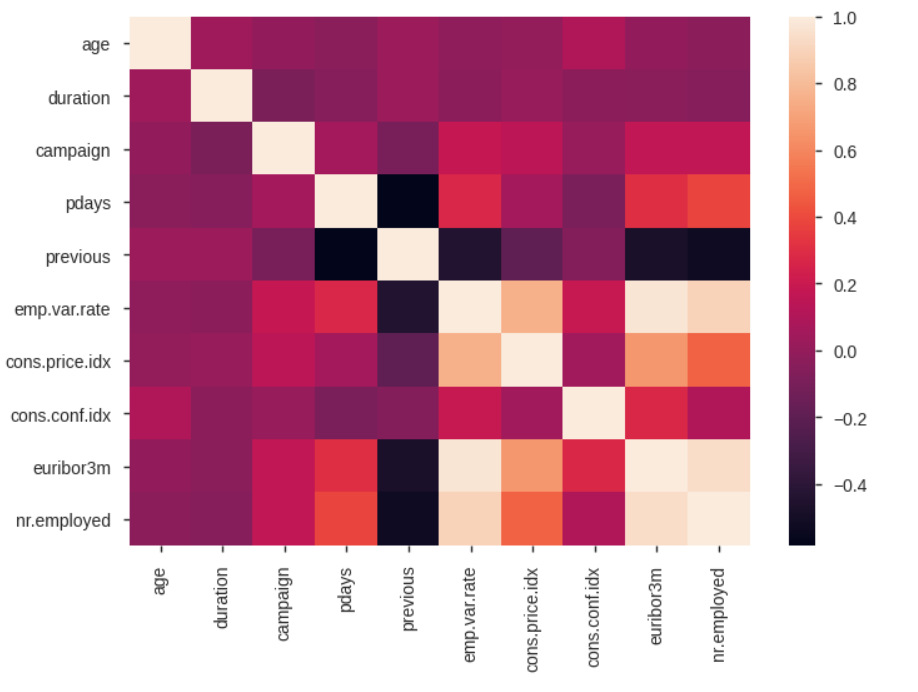
* K-fold cross validation: while training.
* Precision: We care the most about True Positives in this case (Subscribed customers). If we misclassify customers who wouldn’t subscribe as if they would (False Positives), it would be costing money to the bank. Thus, Precision would be the most important metrics among all for checking prediction accuracy.
* ROC curve and AUC(area under curve): For comparisons between different models.
* Recall and F1 scores: To assist Precision and get an overall assessment of the models.

**EDA observations:**

* The average age of customers who bought the term deposit is higher than that of the customers who didn’t.



* The pdays (days since the customer was last contacted) is understandably lower for the customers who bought it. The lower the pdays, the better the memory of the last call and hence the better chances of a sale.
* Surprisingly, campaigns (number of contacts or calls made during the current campaign) are lower for customers who bought the term deposit.

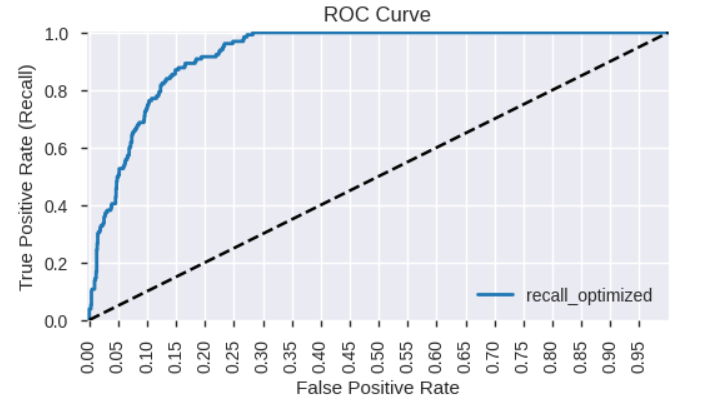


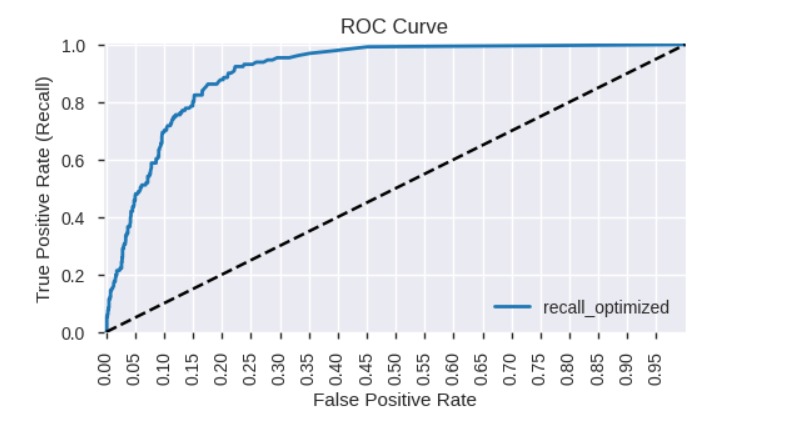
**Steps taken:**

1. Basic Exploratory Analysis
2. identifying relationships (between Y & numerical independent variables by comparing means)
3. Handling categorical features
4. Feature Selection
5. Dropping columns based on data audit report
6. Variable reduction using WOE or log(odds)
7. Variable Reduction using univariate Regression (short list based on Somer's D values)
8. Variable Reduction using Recursive Feature Elimination
9. Variable reduction using Select K-Best technique
10. Variance Inflation Factor assessment
11. Final list of variable selected for the model building from above steps
12. Cross Validation
13. Decision Trees
14. Random Forest
15. XGBOOST
16. Support Vector Machines (SVC)
17. Artificial Neural Networks (ANN)
18. Saving as pickle object
19. Test data for API

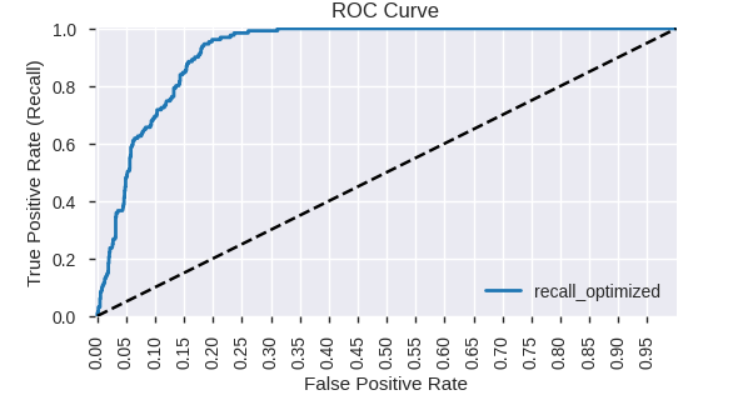
**Model comparison:**

**LR: (92.76%)**



**RF: (Test AUC: 90.74%)**

**XGB: (Test AUC: 92.48%)**



**After tuning:**

* Decision Trees (AUC: 0.9086624775583484)
* Random Forest (AUC: 0.6995958688819038)
* Support Vector Machines (SVC) (AUC:0.7691064211944321)
* Artificial Neural Networks (ANN) (AUC:0.7178335808780354)
* XG Boost (AUC:0.9312389900176159)

**End model deployed**: XG Boost after selecting best features.

**Deployment**: <https://td-subscription-prediction-api.herokuapp.com/>

**References:**

**[1]**7 Ways Machine Learning Can Enhance Marketing. Digital Marketing Institute. (2020). Retrieved 2 November 2020, from <https://digitalmarketinginstitute.com/blog/7-ways-machine-learning-can-enhance-marketing>.

**[2]** UCI Machine Learning Repository: Bank Marketing Data Set. Archive.ics.uci.edu. (2020). Retrieved 2 November 2020, from <https://archive.ics.uci.edu/ml/datasets/Bank+Marketing>.

**[3]** Testers guide for testing machine learning models. Medium. (2020). Retrieved 2 November 2020, from <https://medium.com/analytics-vidhya/testers-guide-for-testing-machine-learning-models-e7e5cea81264>.