Data Science Challenge

Table of Contents

By Harish Damera

Problem Statement:

Building a model to optimize the marketing efforts of an insurance company, by predicting whether customer will respond to the marketing campaign and purchase a policy or not.

Describe your model and why did you choose this model over other types of models?

➢ Random Forest

Random Forest is a tree based model with relatively good accuracy, robust and explanatory.

- They can handle many categorical features and many levels within.
- Most important features for better predictions are provided.
- They don't overfit and mostly have lower classification error.
- Performs well even with unbalanced data sets, as weightages can be adjusted.
- Can discover more complex dependencies and they don't expect linear features.

➢ Confusion Matrix

		PREDICTED	
		No	Yes
ACTUAL	No	764	550
	Yes	25	143

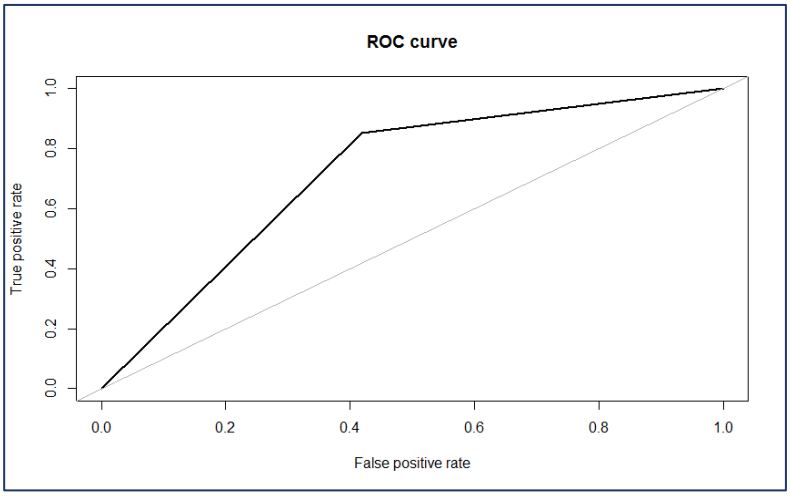
- Assuming that main objective here is to predict potential customers, rather than classifying customers and non-customers accurately.
- It is assumed that the cost associated with email marketing is very less and profit by predicting one extra potential customer is more.
- This model is tuned to achieve less False Negatives and more True Positives.

➢ Performance

Accuracy	0.61
Recall	0.85

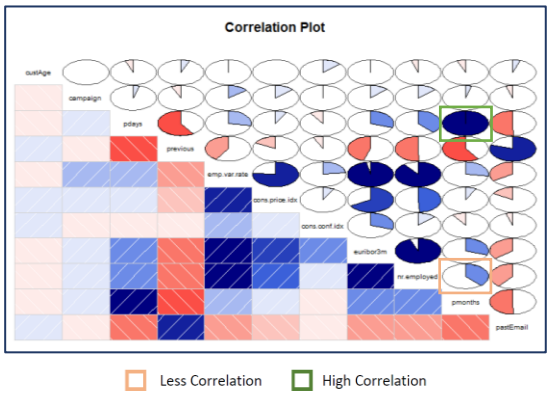
- Model might have low accuracy but has high predictive power.
- The goal is to optimize the marketing efforts and have fewer False Negatives.
- Trade-off between accuracy and recall is needed for optimizing.
- Predicting 85% potential customers and eliminating more than 60% non-customers.
- Other models had far better accuracy but less predictive power.

➢ ROC Curve



➢ Feature Selection:

- Correlation: Eliminated attributes with high correlations considering variable importance list from baseline Random Forest model. Pdays and emp.var.rate are second important variables among correlated, hence removed.
- Chi-square test: Performed to check for any Independence. p-values of housing, loan and day_of_week are greater than 0.05 significance level.



➢ Distributions: Box Plots (Outlier Detection and Data Distribution)

➢ Imputation: K-Nearest Neighbors (KNN)

➢ Data Partition: implemented Stratified Sampling to equally consider samples of both classes.

Describe any other models you have tried and why do you think this model performs better?

Implemented Logistic Regression and c5.0 boosting algorithms. Random Forest gave best recall value and is able to predict potential customers more than any other models. It predicted least False Negatives and more True positives, supporting our assumption.

➢ Logistic Regression

- Logistic regression performs well with fewer categorical variables and fewer levels within.
- Even though it is easy to compute and interpret, multicollinearity might exist.
- Predictions are not great. Higher False Negatives compared to Random Forest.

		PREDICTED	
		No	Yes
ACTUAL	No	1135	179
	Yes	63	105

➢ c5.0 Boosting

- Boosting trials set to maximum i.e., 100
- Model is able to differentiate only True Negatives. Recall is very low.
- Applied Cross Validation and best Recall iteration is used for model building.
- No great improvement in the predictions. Still low Recall compared to Random Forest.

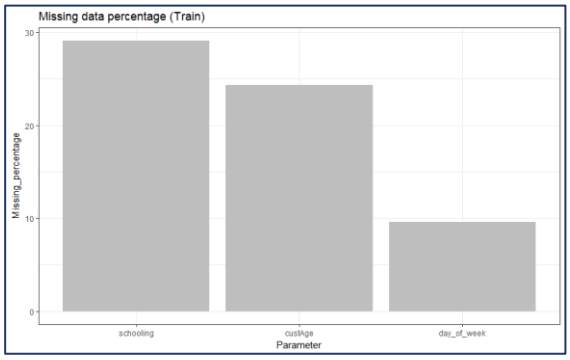
		PREDICTED	
		No	Yes
ACTUAL	No	1242	72
	Yes	75	93

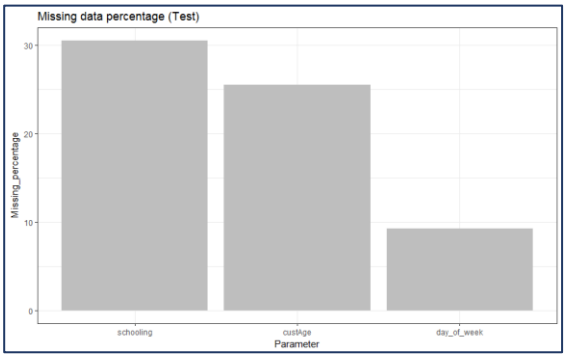
		PREDICTED	
		No	Yes
ACTUAL	No	1241	73
	Yes	78	90

How did you handle missing data?

Applied **KNN (K nearest Neighbors)** algorithm to impute the missing data. Missing values are imputed using other attributes that are more like this attribute. Similarity is determined using a distance function.

- I choose to use nearest 5 neighbors for imputation.
- Imputation by KNN gives better chances of achieving accurate models when compared to mean/mode imputation, for this data.





How did you handle categorical (string) data?

➢ Chi-squared Test

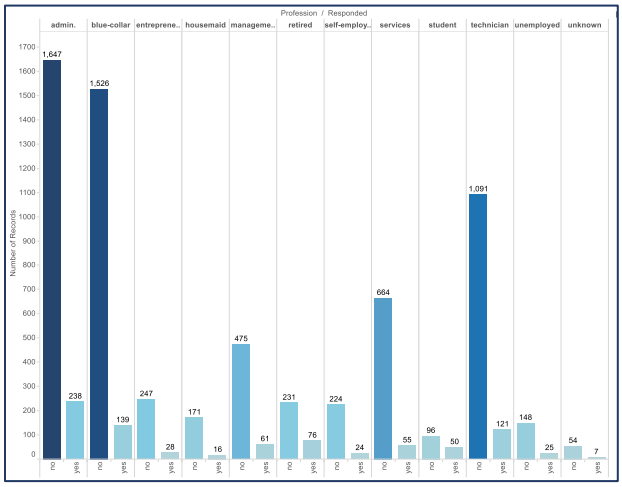
- Chi-squared Test is performed to check for any Independence at 0.05 significance level.
- p-values for housing, loan and day_of_week (Monday and other) are greater than 0.05. Here we do not reject the null hypothesis that these attributes are independent with the target variable “responded”. Hence, not considered for further analysis.

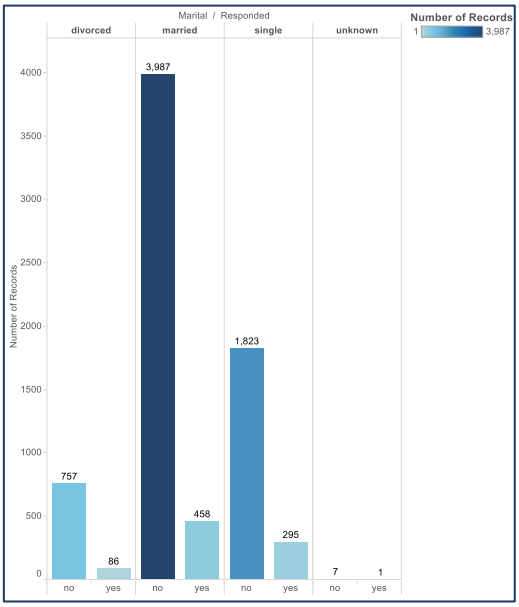
➢ Proportion of respondants per each category (Univariate Analysis)

Categories with similar proportion of respondants are consolidated into single category.

- Unemployed, unknown, management and admin. categories of Profession attribute are consolidated into a single “admin.” category. blue-collar, housemaid, services, self-employed, entrepreneur and technician are consolidated into single “blue-collar” category.
- basic.4y, basic.6y, basic.9y, high.school and professional.course categories of schooling attribute are consolidated into a single “high.school” category. “Illiterate” category is merged into “unknown” category.
- sep, oct, mar and dec months of month attribute are consolidated into “dec” category. Aug, jul, jun, may and nov months are consolidated into “jun” category
- As the number of instances in unknown category of “Marital” attribute are considerably low, it is merged into majority category “married”.
- Default “Yes” had only 1 instance, and is merged into “unknown” category.
- “Unknown” category of loan attribute is merged into “no” category.

➢ Below bar plots are done in Tableau



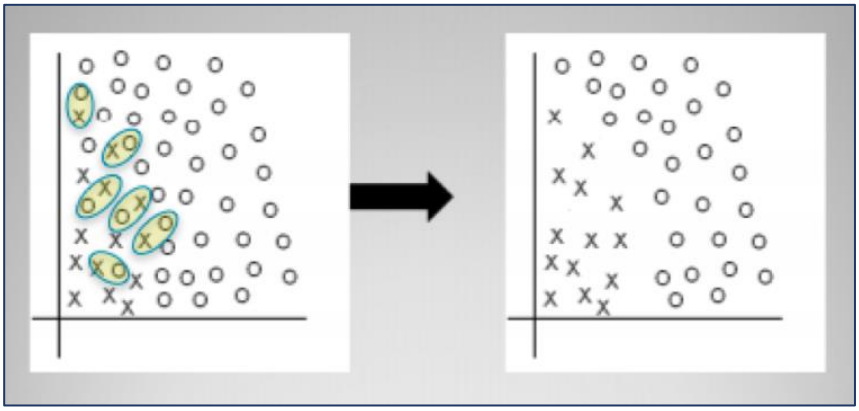


How did you handle unbalanced data?

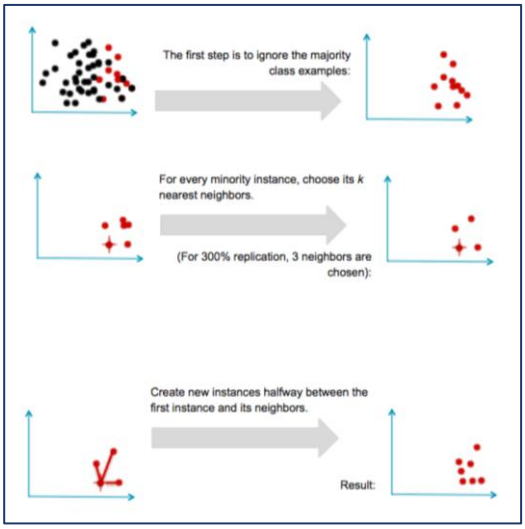
- Here minority class samples are much more important and hence needed to be classified with much attention.
- TOMEK Links and SMOTE algorithms are applied to overcome this unbalanced data issue. Initially minority class had 672 instances.

➢ TOMEK LINK

- Tried to establish a border between minority and majority classes, making the minority region more distinct by Tomek's algorithm.
- Tomek links are the pairs of opposing class data points that are very close together. This algorithm looks for such pairs and removes the majority instance of that pair.



- SMOTE (Synthetic Minority Over-Sampling Technique)
    - Training data is balanced by synthesizing new minority class samples by SMOTE algorithm.
    - This method will only fill in the convex hull of existing minority instances, but will not create new exterior regions of minority instances.
    - New minority instances are created by interpolating between existing instances. I choose to increase the minority class by 400%.



How did you test your model?

Marketing_training dataset is partitioned into train and test datasets by stratified sampling technique for internal Evaluations (80:20 ratio). Models are built on “Smoted” data and verified on “test” data. Multiple metrics are used to evaluate the performance.

- Cross Validation
    - 5-fold cross validation is applied to test model performance.
    - Mean Accuracy 74% and Mean Recall is 95%. Max Recall is achieved in 3rd iteration.
- Confusion Matrix
    - Model predicted less False Negatives (25) and more True Positives (143)
- Accuracy & Recall
    - Predicting 85% potential customers and eliminating more than 60% non-customers.
- MCC (Matthews correlation coefficient)
    - Measures the quality of binary classifications. Numbers close to +1 indicates good predictions and close to -1 indicates worse predictions.
    - Here MCC = 0.27, because of tradeoff between Accuracy and Recall.
- ROC
    - Area under the curve (AUC): 0.716