

Capstone Project 1: Final Report

1. Problem statement

Customer satisfaction is an important factor for businesses. Every business wants to know happy and unhappy customers to take action. We will predict Santander Bank customers' happiness. So they can take action before they leave the bank.

2. Dataset

I will be using Kaggle project data.

The source link:

[Santander Customer Satisfaction](#)

This data set has 370 anonymous features and 76020 rows. Target feature is binary, happy or unhappy customers. The goal is to predict customers' satisfaction.

In the TARGET feature 0 represents happy and 1 represents unhappy customers.

The first thing I check is the percentage of happy and unhappy customers. 96/4 ratio means this dataset is unbalanced. I have to consider under and over sampling methods.

Secondly the missing values are checked. There are no missing values in the dataset.

3. Feature Selection

All features are anonymous so we have to use statistical techniques to see whether the features are important or not.

The original number of features : **370**

- **Basic Methods**

Remove constant and quasi-constant features 370 -> 273

First I applied basic methods and removed constant and quasi-constant features. With the help of this method I reduced the number of features from 370 to 273.

- **Univariate Selection Methods 273 -> 30 -> 22**

Secondly I applied univariate methods like ANOVA F-value and Fisher. By applying these two methods I choose the top 30 features. Then I looked at the common features of both methods. At the end 22 features are found in common.

- **Correlation Matrix Method 22 -> 11**

I created a correlation matrix and found features with a correlation greater than 0.90. I drop those features and 11 features are left.

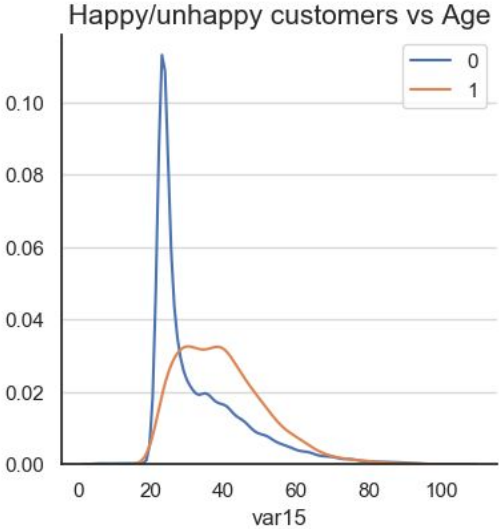
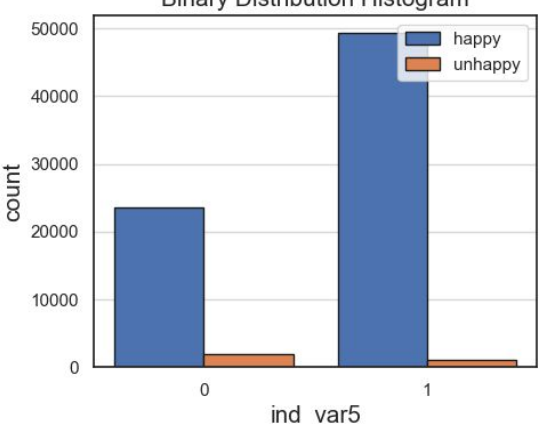
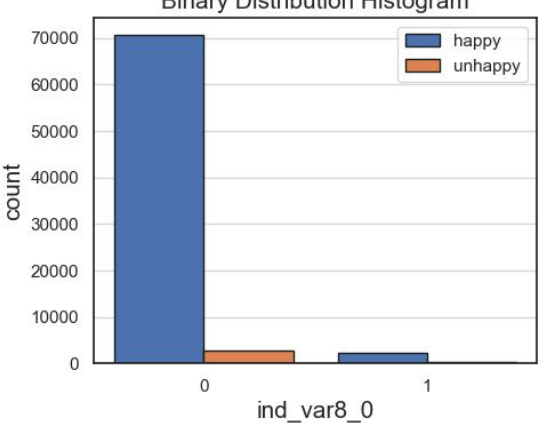
I did EDA and indeed these features are helpful features to explain my target.

4. Exploratory Data Analysis (EDA)

Selected features are:

1. var15',
2. 'ind_var5',
3. 'ind_var8_0',
4. 'ind_var12_0',
5. 'ind_var13_0',
6. 'ind_var30',
7. 'num_var30_0',
8. 'num_var30',
9. 'num_var42',
10. 'saldo_var30',
11. 'var36'

Let's analyze them visually and statistically.

	<p>Fig 1: var15</p> <p>has definitely resolving power. According to a Kaggle form post var15 is the age of the customers. Younger people are mostly happy customers.</p>
	<p>Fig 2: ind_var5</p> <p>The customer dissatisfaction is 3.3 times more likely when ind_var5 is absent obtained from $P(\text{Target}=1 \text{ind_var5}=0) / P(\text{Target}=1 \text{ind_var5}=1) = 7.7/2.1 = 3.7$</p>
	<p>Fig 3: Ind_var8_0</p> <p>The customer dissatisfaction is 0.4 times more likely when ind_var8_0 is absent (obtained from $P(\text{Target}=1 \text{ind_var8_0}=0) / P(\text{Target}=1 \text{ind_var8_0}=1) = 0.4$)</p>

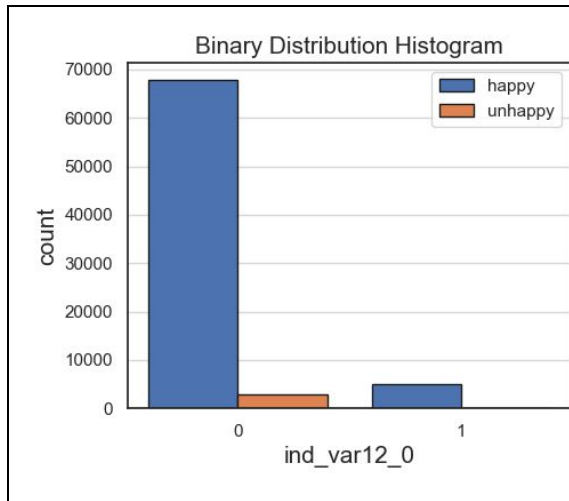


Fig 4: ind_var12_0

The customer dissatisfaction is 3.5 times more likely when ind_var12_0 is absent
 obtained from $P(\text{Target}=1|\text{ind_var12_0}=0) / P(\text{Target}=1|\text{ind_var12_0}=1) = 3.5$

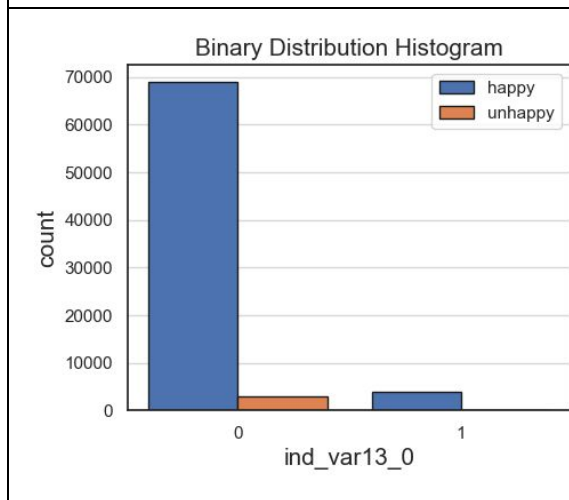


Fig 5: ind_var13_0

The customer dissatisfaction is 6.1 times more likely when ind_var13_0 is absent

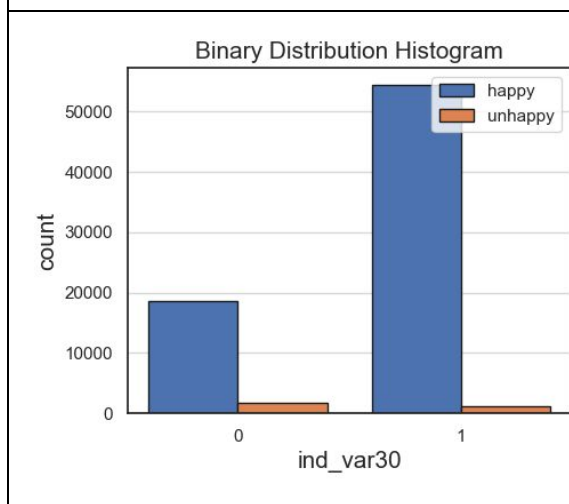


Fig 6: ind_var30

The customer dissatisfaction is 4 times more likely when ind_var30 is absent

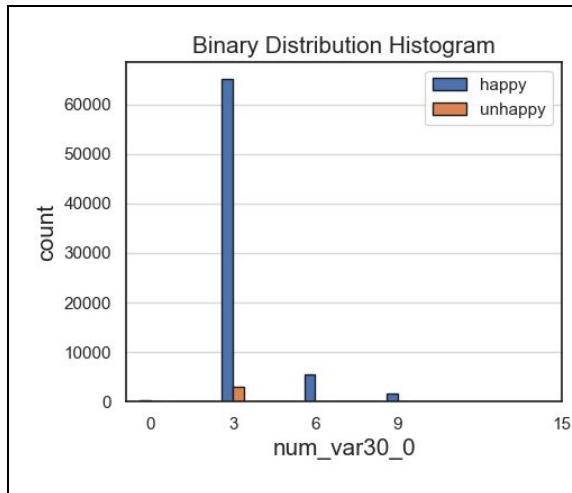


Fig 7: num_var30_0

The customer dissatisfaction is more likely when num_var30_0 is 3,6, or 9.

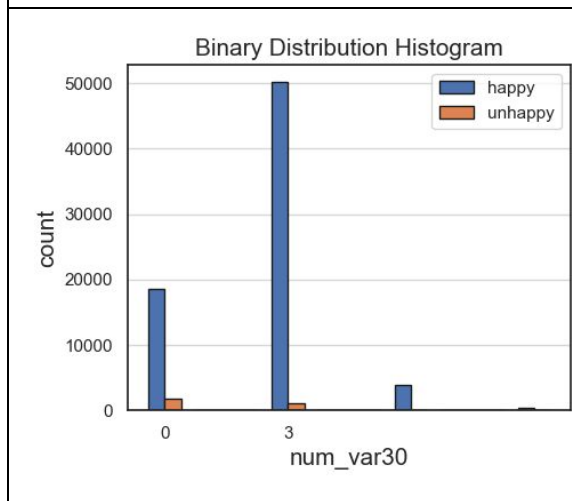


Fig 8: num_var30

The customer dissatisfaction is more likely when num_var30_0 is 0.

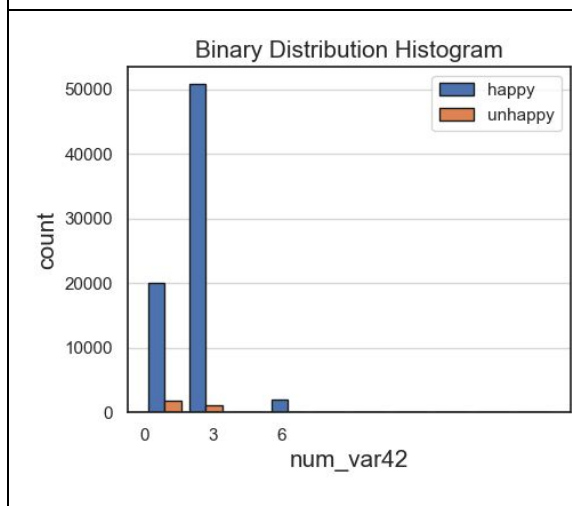
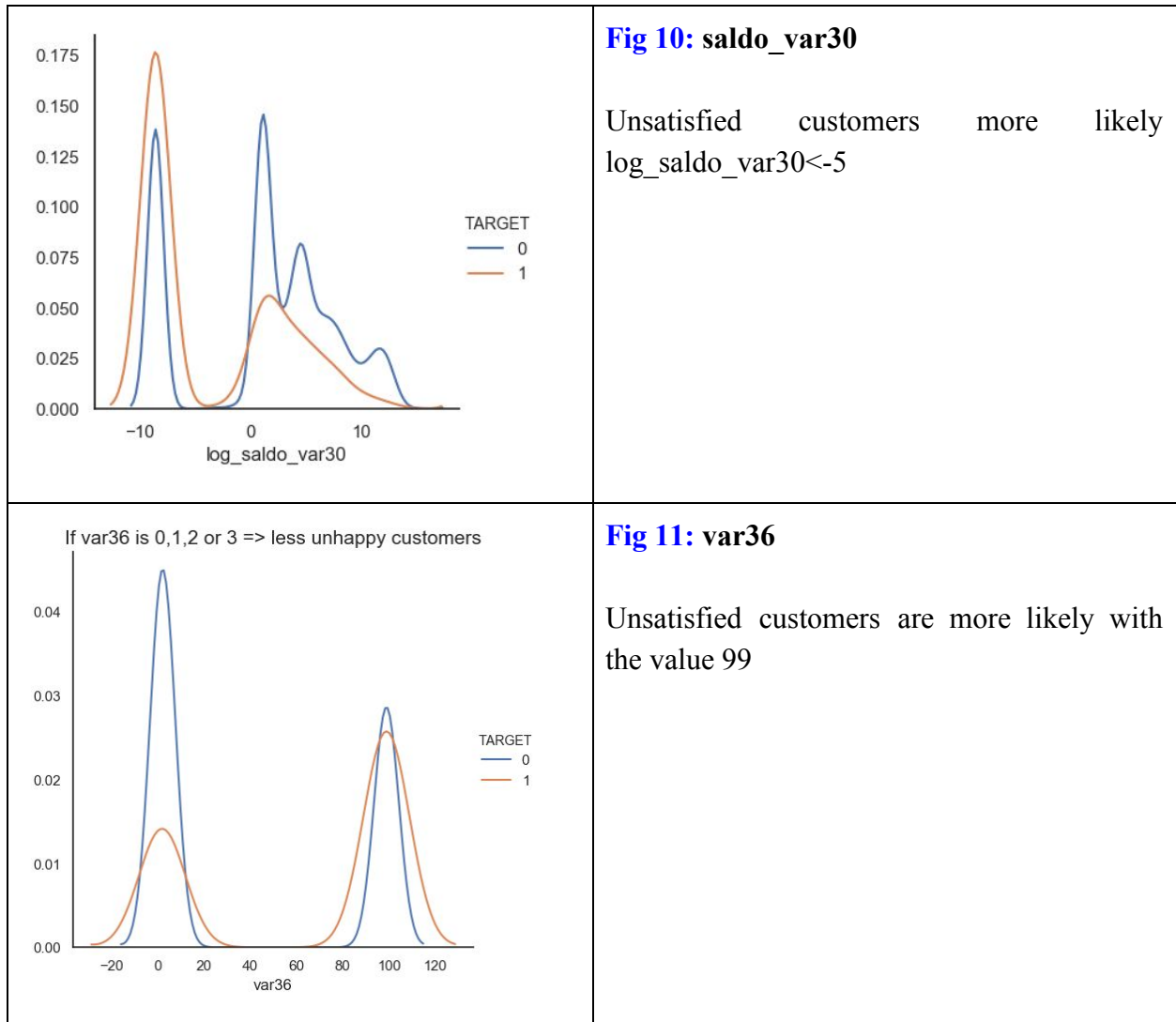


Fig 9: num_var42

The customer dissatisfaction is more likely when num_var42 is 0.



5. Predictive Models

I created a new dataframe(X4) with these selected features and target column. Target is labeled and binary so I need to use a supervised predictive model. Five predictive models are applied: logistic regression, support vector machine, random forest, k nearest neighbor, and decision tree classifiers.

This data set is unbalanced, this should be considered during model training.

First I split the data set into train(X_train, y_train) and test(y_train, y_test) sets with test size = 0.2. (Random state=42). On the train set I did model training. After training I reported confusion matrix and classification report. These are my base scores. Confusion matrix is also plotted.

For each model I did over sampling (smote) and random under sampling. Five main predictive models with their base, smote , and random under sampling scores, I got totally 15 different reports and scores. I compared weighted average f1-scores, recalls for 1, and AUC.

I created 3 tables where model names are in the rows and base, smote and random under sampling models go to columns. Scores fill the tables.

Model Name	Base Model	Smote Model	Random Under Sampling Model(RUS)
LR	0.94	0.77	0.77
SVM	0.94	0.25	0.21
RF	0.94	0.86	0.83
kNN	0.94	0.89	0.87
DT	0.94	0.79	0.86

Table 1:Weighted average f1-scores

Model Name	Base Model	Smote Model	Random Under Sampling Model(RUS)
LR	0	0.68	0.67
SVM	0	0.96	0.97
RF	0.02	0.59	0.71
kNN	0.04	0.23	0.32
DT	0	0.78	0.69

Table 2: Recall - 1

Model Name	Base Model	Smote Model	Random Under Sampling Model(RUS)
------------	------------	-------------	----------------------------------

LR	0.74	0.76	0.76
SVM	0.60	0.60	0.71
RF	0.74	0.74	0.79
kNN	0.64	0.62	0.58
DT	0.82	0.79	0.81

Table 3:AUC

I decided the decision tree is the best.

5.1. Parameter Tuning for DT

After deciding for the best model I did parameter tuning. Grid search is used for parameter tuning. To make comparison and understand better grid search is performed on base, smote, and under sampling models. After finding the best parameters, the train dataset is trained with those best parameters.

At the end, I get scores using my test data set.

My reported scores are at Table 4.

Model Name	Weighted average F1	Recall 1	AUC
DT Base	0.94	0.03	0.60
DT Base with best parameters	0.94	0.03	0.81
DT Smote	0.86	0.58	0.69
DT Smote with best parameters	0.85	0.66	0.79
DT RUS	0.82	0.68	0.70
DT RUS with best parameters	0.85	0.71	0.85

Table 4: Scores for different DT models

6. Conclusion

I recognized that when I performed parameter tuning, recall for 1 and AUC scores are always increased. AUC is always the same for untuned models(~0.69). Sampling models increased the recall 1 score.

The best AUC score 0.85 is coming from tuned DT - RUS model and also this model gives the best recall -1 score, 0.71, over all.