FIN CORP BANK

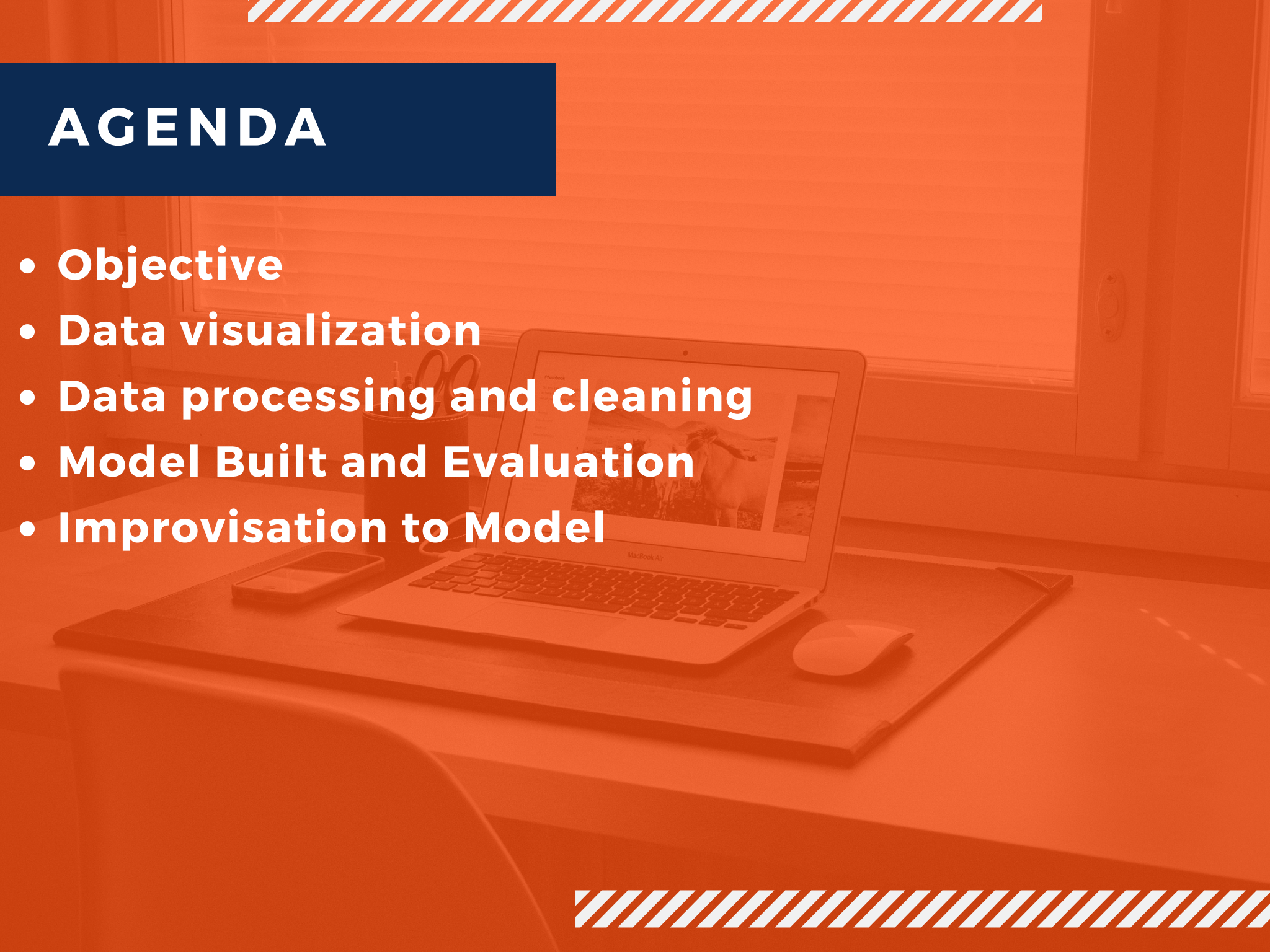
PREDICTING CUSTOMER CHURN (SATISFACTION)

2017


FINCORP
SWAZILAND DEVELOPMENT FINANCE CORPORATION

AGENDA

- Objective
- Data visualization
- Data processing and cleaning
- Model Built and Evaluation
- Improvisation to Model



1 OBJECTIVE & DATA PREPROCESSING

Objectives of current project is:

1) Predicting "class" (1: High energy seismic bump occured) ((target variable V20) is a classification problem

2) The given data set has following attributes (19805 rows and 371 features, 67 values have 0 variance

- Categorical Variables - 61 # 61 vars have values b/n 0, 1, # 34 vars have values b/n 0, 3
- Numerical Variables - 200

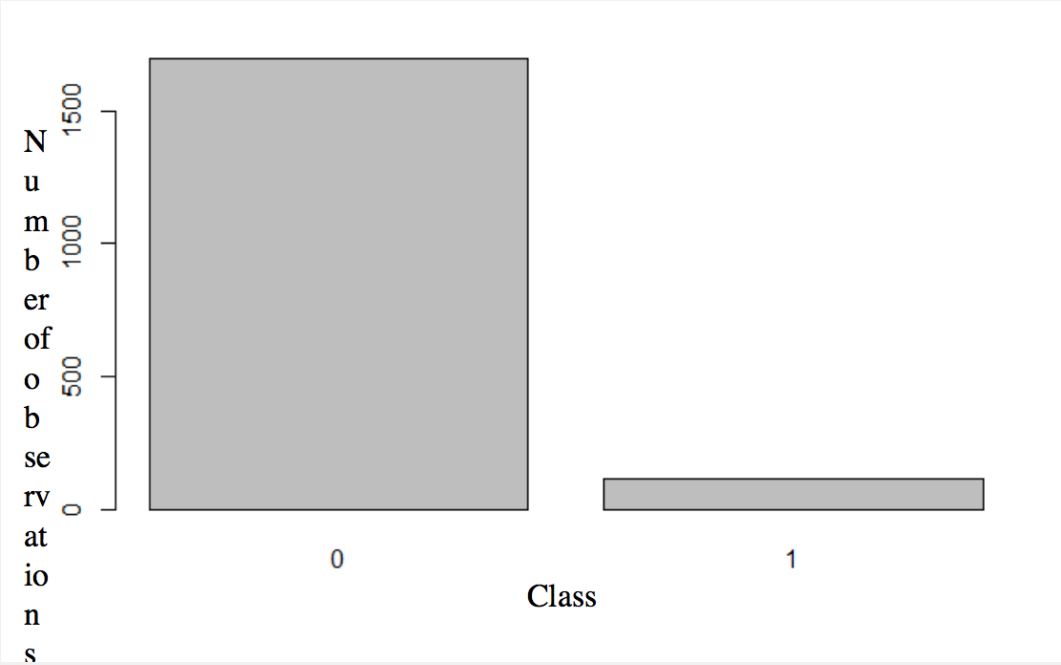
3) 6339 train, 2537 Validation, 3961 test records

1 OBJECTIVE & DATA PREPROCESSING

1) There are 0 missing values Zero

2) Heavy Class imbalance. Used "Smote" on train to prevent class imbalance # 10.04% to 40%

3) Scaled all features



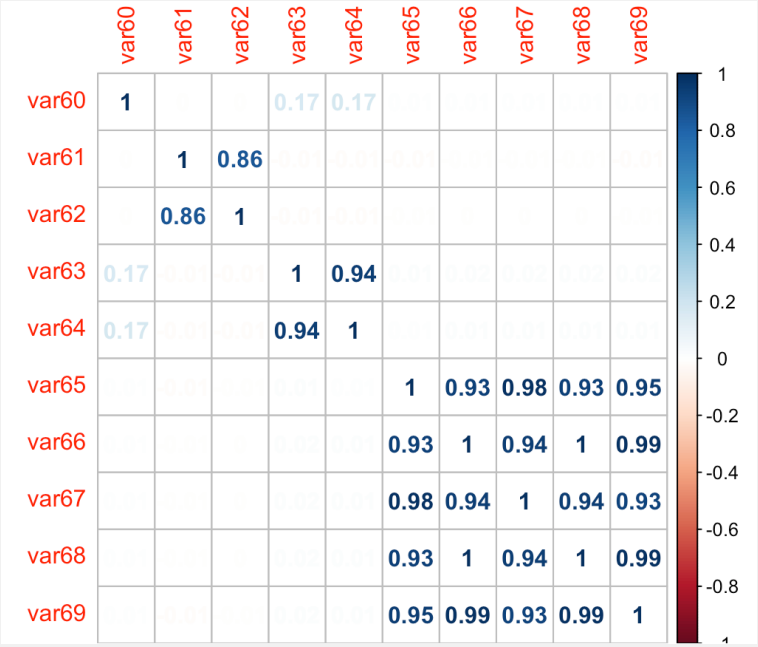
2

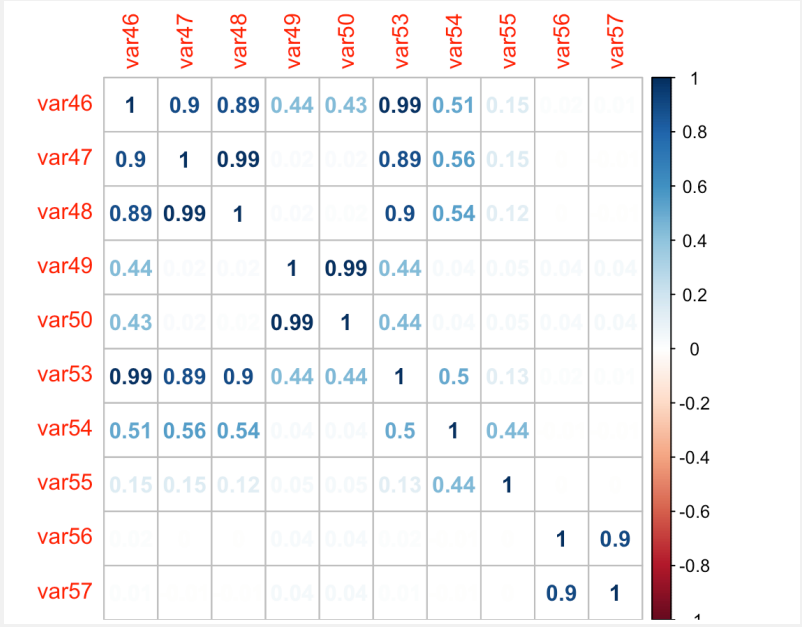
SEISMIC DATAVISUALIZATION - CORRELATION BETWEEN FACTORS

Some trends!

#var 49, var 50 have heavy correlation

var 47, var 49 have heavy correlation





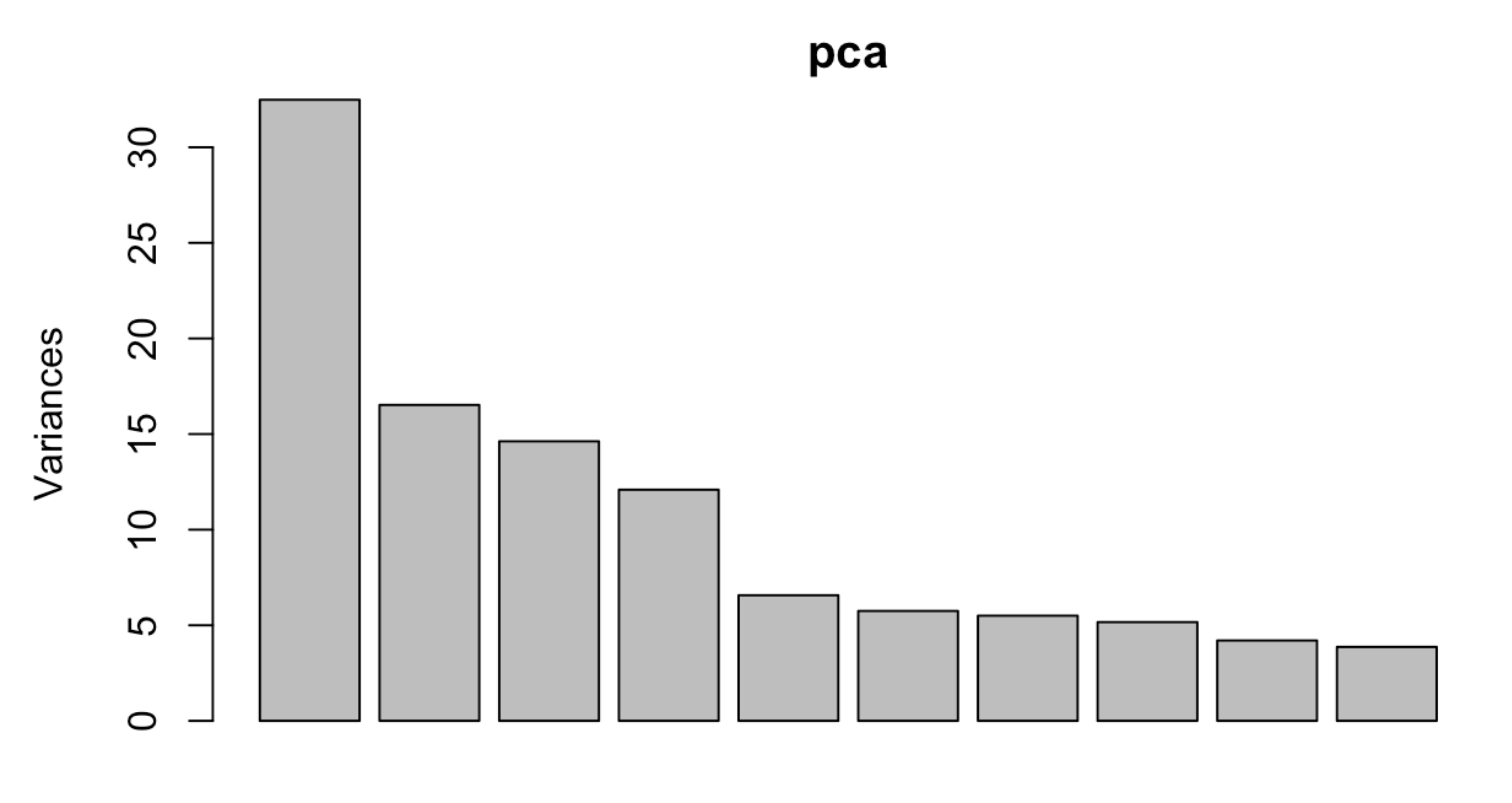
3 DATA PRE PROCESSING

Removed Near Zero Variance and conducted PCA. Retained 15 transformed features

```
> pca <- prcomp(train_rnzv[,!(names(train_rnzv) == "TARGET")], center = T, scale. = T)
> summary(pca)
Importance of components:
                          PC1    PC2     PC3    PC4    PC5     PC6     PC7     PC8
Standard deviation     5.6996 4.0648 3.82372 3.4771 2.5631 2.39748 2.34536 2.27257
Proportion of Variance 0.2166 0.1101 0.09747 0.0806 0.0438 0.03832 0.03667 0.03443
Cumulative Proportion  0.2166 0.3267 0.42419 0.5048 0.5486 0.58691 0.62358 0.65801
                           PC9    PC10    PC11    PC12    PC13    PC14    PC15
Standard deviation     2.05021 1.96627 1.88582 1.78586 1.74281 1.57099 1.51073
Proportion of Variance 0.02802 0.02577 0.02371 0.02126 0.02025 0.01645 0.01522
Cumulative Proportion  0.68603 0.71180 0.73551 0.75678 0.77702 0.79348 0.80869
                          PC16    PC17   PC18    PC19    PC20    PC21    PC22
Standard deviation     1.45195 1.42575 1.3745 1.26762 1.24542 1.19312 1.12843
Proportion of Variance 0.01405 0.01355 0.0126 0.01071 0.01034 0.00949 0.00849
Cumulative Proportion  0.82275 0.83630 0.8489 0.85961 0.86995 0.87944 0.88793
```

3 DATA PRE PROCESSING

Removed Near Zero Variance and conducted PCA. Retained 15 transformed features accounted for over 80 variance



4 MODEL BUILT AND EVALUATION

Typical Accuracy : 81%

AUC value is 0.655

4 MODEL BUILT AND EVALUATION

- We realized that for this seismic domain the AUC is the important aspect. Tried the below models in the stipulated time. KNN along with GBM provided best results

Algorithms	Val Accuracy	Test Accuracy	AUC
Decision Tree	0.8564	0.8634	0.65
KNN (k=1)	0.8868	0.804	***0.67***
Random Forest	0.8699	0.847	0.6355
Bagged DT with tuning	0.8786	0.8687	0.63
GBM	0.9093	0.8705	0.65
SVM - with tuning	0.88	0.85	0.634

5 IMPROVISATIONS TO MODELS WITH TIME

Novel methods to Feature engineering viz. splitting looking at the patterns, binning few of the features etc into one single bin

The Domain knowledge would have made more accurate models

6 CHALLENGES

Understanding the precise definition of the variables, might have accomplished feature engineering

Large data set and larger set of features

'R crashed when the SVM, GBM

IF YOU HAVE QUESTIONS, DS!