Prediction of Customer Attrition in the Telecom Industry using Machine Learning

Anish Mahapatra | 31 May, 2021

LIVERPOOL JOHN MOORES UNIVERSITY

Agenda

01 Introduction

02 Literature Review

03 Aim and Objective

04 Research Methodology

05 Results and Discussions

06 Conclusions and Future Recommendations

LIVERPOOL JOHN MOORES UNIVERSITY

Introduction

The telecom industry is valued at **$1658 billion**.

The percentage of customers a telecom operator can retain decides the profits of the company.

The Cost of Customer Acquisition

Acquiring new customers is **5-10 times** more expensive than keeping existing customers loyal

The Cost of Customer Churn

The cost of customer churn is **$10 billion** globally every year

The Average churn rate of customers

Companies, on average lost **10-30%** of their customers annually

The Profits from Customer Retention

If customer retention cost was increased by 5%, *profits* would increase by **50-75%**

LIVERPOOL JOHN MOORES UNIVERSITY

Literature Review

Exploratory Data Analysis

Visual analysis, univariate analysis, bivariate analysis,

Modelling

Single machine learning models, meta-heuristic models, hybrid models, data mining techniques

Feature Selection

Mostly XGBoost was used to conduct feature selection

Evaluation Metrics

The models were evaluated using accuracy or AUC scores are the primary methods of assessment

Feature Engineering

Box Cox Transformation, Class Balancing, Handling Categorical variables, Standardization, Normalization

Gap: Interpretable Machine Learning

There was a gap of research that had good results and performed interpretable machine learning

Aim and Objective

Aim

The aim of the paper is to develop a **trustworthy and interpretable model** that will predict customers that will churn.

Objectives

- ***Visualize patterns*** of customer behavior
- ***Feature Selection*** to identify important attributes
- Implement ***class balancing*** techniques to improve model performance
- Develop and ***evaluate machine learning models***
- To help the business make sense of predictions, leverage ***interpretable machine learning***

LIVERPOOL JOHN MOORES UNIVERSITY

Research Methodology

01 Exploratory Data Analysis

Data Understanding, Distribution of variables, Missing Value Analysis, Outlier Analysis, Bivariate Analysis

02 Statistical Tests

Chi-Square Test, ANOVA, Probability Distribution using Kernel Density Estimation

03 Feature Engineering

One-Hot Encoding, Feature Importance Analysis, Standardization (After train - validation - test split), Class Balancing

04 Model Building

Train-test split, Baseline models, Hyperparameter tuning, Class Balancing - Oversampling using SMOTE-NC,

05 Model Evaluation

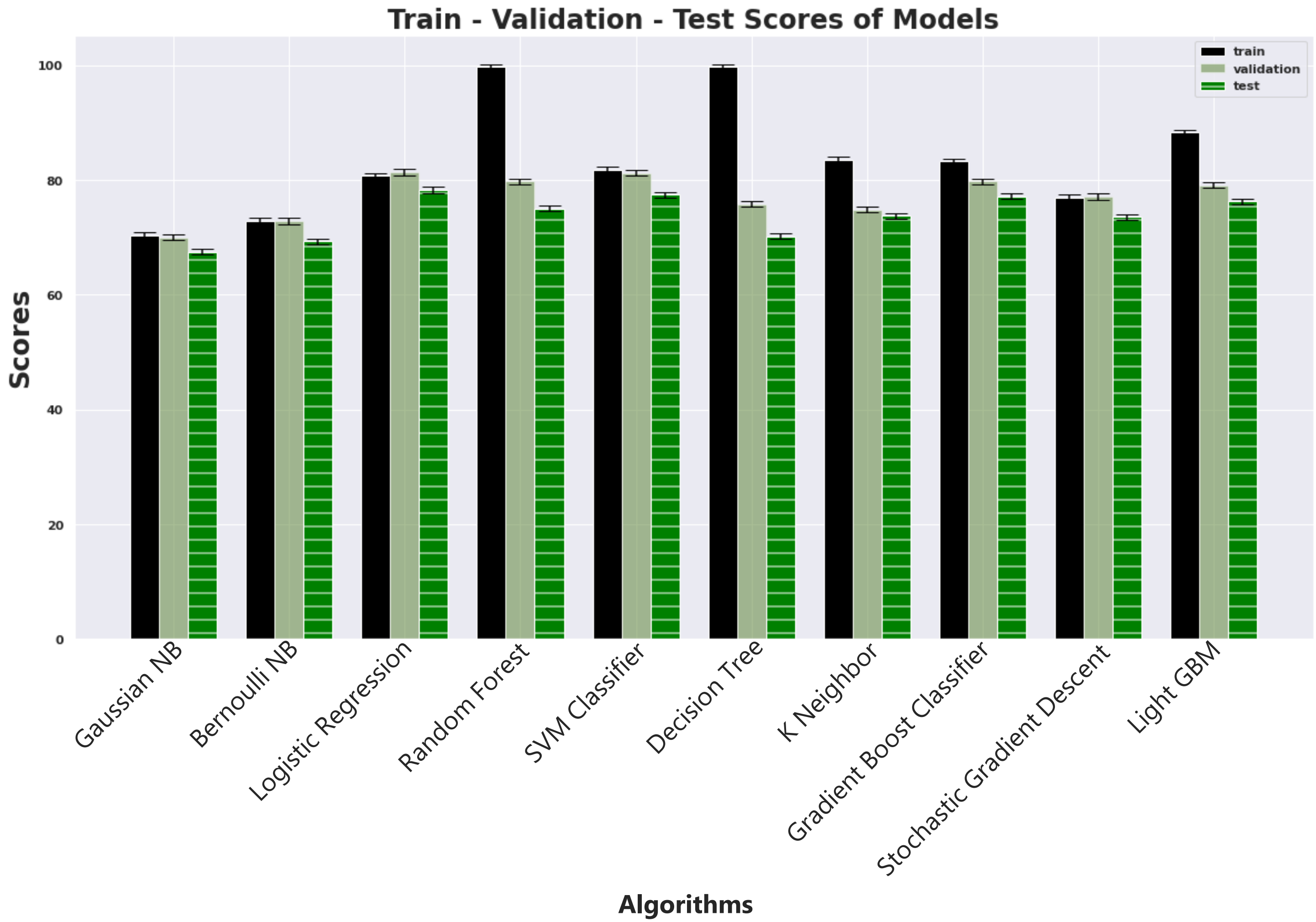
Model Evaluation on Train-Validation-Test Data, after Class Balancing, and after Oversampling using SMOTE-NC

06 Model Interpretability

Model Interpretability using Locally Interpretable Model - Agnostic Explanation (LIME) and Shapely Additive Explanations (SHAP)

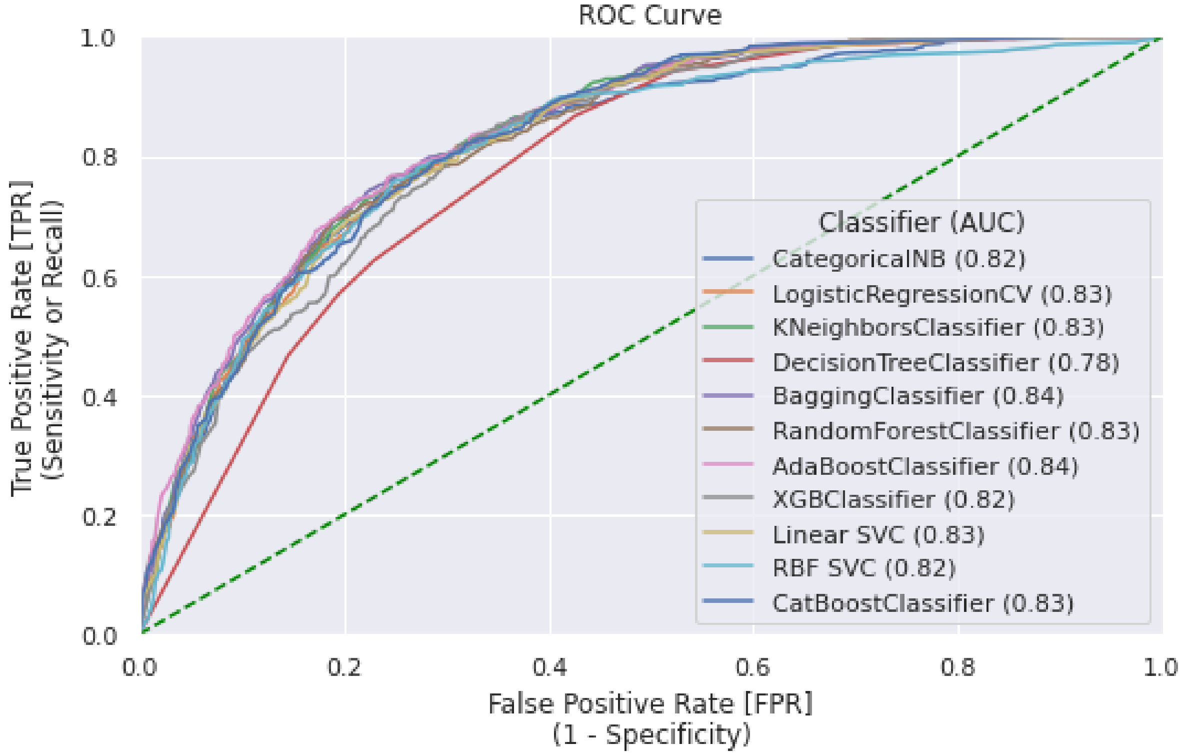
LIVERPOOL JOHN MOORES UNIVERSITY

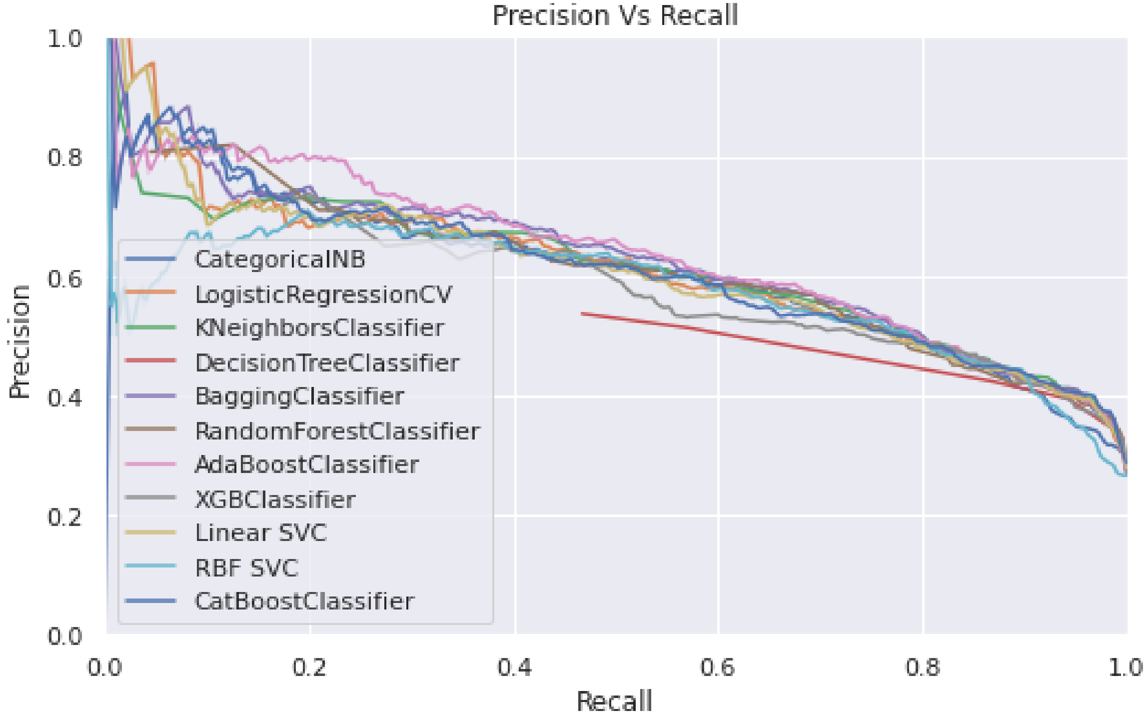
Results | Train-Validation-Test



- The data has been split into train, test and validation
- Logistic regression has the highest accuracy on the test data of 78.30%
- Bernoulli Naïve Bayes has the highest AUC Score of 0.74
- Random Forest and Decision Tree are overfit on the train data with accuracies of 99.75%
- Support Vector Machine and Gaussian Naïve Bayes are the other models that performed well

Results | Model Performance

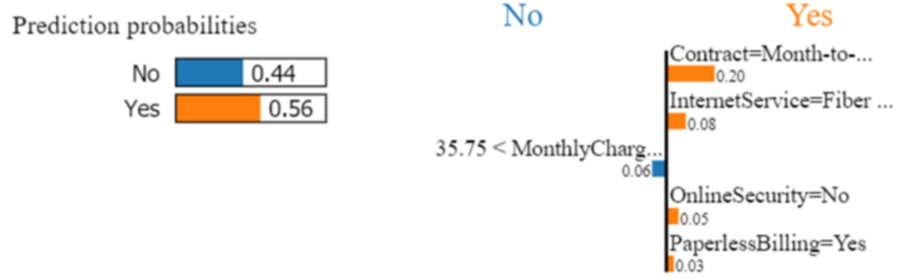




- The results are obtained after **class balancing** using SMOTE-NC and **hyperparameter tuning** using Randomized Search CV
- The Decision Tree **AdaBoost** Classifier and **Bagging** Classifier have the *highest AUC score* of 0.84
- All models (except one) have *AUC scores of greater than 0.80* and test accuracy scores of greater than 70%
- **XGBoost** and **CatBoost** are the ensemble models that have the highest accuracy of about 76%

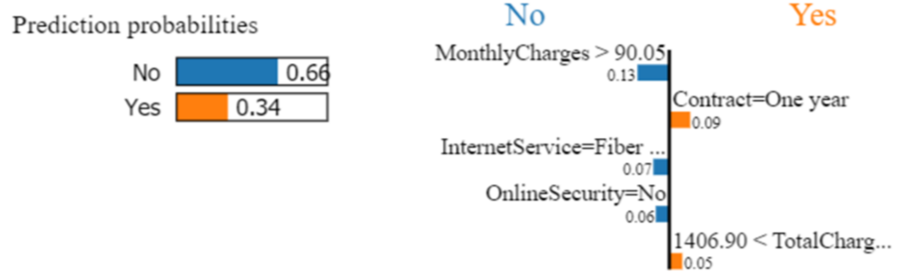
Results | Interpretable Machine Learning

```
Document id: 40
Probability(0) = 0.44290692
Probability(1) = 0.5570931
True class: 0
```



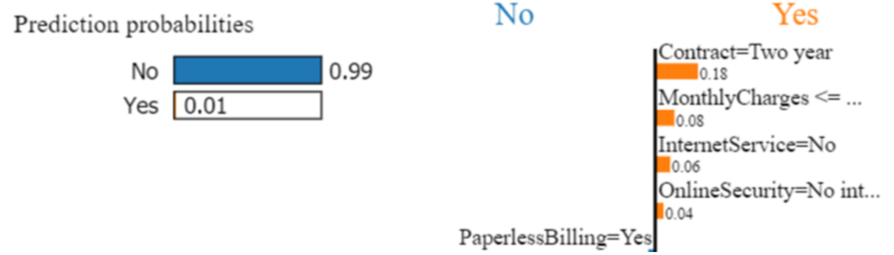
Feature	Value
Contract=Month-to-month	True
InternetService=Fiber optic	True
MonthlyCharges	68.60
OnlineSecurity=No	True
PaperlessBilling=Yes	True

```
Document id: 51
Probability(0) = 0.6638942
Probability(1) = 0.3361058
True class: 0
```



Feature	Value
MonthlyCharges	96.65
Contract=One year	True
InternetService=Fiber optic	True
OnlineSecurity=No	True
TotalCharges	1588.25

```
Document id: 19
Probability(0) = 0.9875589
Probability(1) = 0.012441074
True class: 0
```



Feature	Value
Contract=Two year	True
MonthlyCharges	19.55
InternetService=No	True
OnlineSecurity=No internet service	True
PaperlessBilling=Yes	True

- LIME stands for Locally Interpretable Model Agnostic Explanation
- The output is a set of explanations that explain each feature's contribution to predicting a data point
- SHAP is used for global interpretation of the features using Shapely values

LIVERPOOL JOHN MOORES UNIVERSITY

Conclusions and Future Recommendations

Conclusions

- **Interpretable machine learning models** such as SHAP and LIME can help the business understand the underlying mechanism of the predictions by the models
- The pipeline that gives the best result includes **class balancing** using *SMOTE-NC*, performing **cross validation** and **hyperparameter tuning** using *Randomized Search CV*
- One of the better models obtained was **CatBoost** with an AUC score of **0.83** and accuracy of **76.45**.
- **Ensemble models** with balanced data tend to give better results as compared to individual machine learning models

Conclusions and Future Recommendations

Future Recommendations

- **Deep Learning models** can be attempted to leverage the interconnections within the data to provide better results
- The suggested pipeline can be attempted in a **real world setting** to check revenue generated
- **More factors** such as demographic information, streaming data and more historical records can be included in the machine learning modelling pipeline to *improve performance*

Prediction of Customer Attrition in the Telecom Industry using Machine Learning

Anish Mahapatra |

Thank you.

LIVERPOOL JOHN MOORES UNIVERSITY