The increasing Cost of Butter: A comparison study between Ireland and France

Author: M. Moran

e-mail: [mmoran@student.cct.ie](mailto:mmoran@student.cct.ie)

Student ID: SBA22220

Continuous assignment No.2 submitted as part of Module 1 assessment for the Degree of Master of Science in Data Analytics

**CCT College Dublin**

**Assessment Cover Page**

*To be provided separately as a word doc for students to include with every submission*

|  |  |
| --- | --- |
| **Module Title:** | Programming for DA  Statistics for Data Analytics  Machine Learning for Data Analysis  Data Preparation & Visualisation |
| **Assessment Title:** | The increasing Cost of Butter: A comparison study between Ireland and France |
| **Lecturer Name:** | Sam Weiss  Marina Iantorno  Muhammad Iqbal  David McQuaid |
| **Student Full Name:** | Michelle Moran |
| **Student Number:** | SBA22220 |
| **Assessment Due Date:** |  |
| **Date of Submission:** |  |

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

List of Contents

Declaration

List of Contents

List of Tables

List of Figures

Abstract

Acknowledgements

Chapter 1: Introduction

1.1 Problem Definition

1.2 Research Objectives

1.3 Target Variable

Chapter 2: Methodology

2.1 Introduction

2.2 Data Sourcing and Preparation

2.3 Data Limitations

2.4 Feature Scaling

2.5 Feature Engineering

2.6 Feature Selection

2.7 Class Imbalance

2.8 Hyperparameter Tuning

2.9 Model Classifiers

2.10 Primary Research

Chapter 3: Results and Discussion

3.1 Introduction

3.2 Evaluation Metrics

3.3 Model Results

3.4 Conclusion

References

Appendices

List of Tables

Table 1 Kendall Rank Correlation Coefficient values and their interpretation. 44

Table 2 Final Data frame post feature selection excluding target variable. 47

Table 3 Standard parameter test set results on 50% of the balanced data set. 71

Table 4 Hyperparameter tuned test set results on the imbalanced data set. 72

Table 5 Hyperparameter tuned test set results on the balanced data set. 73

Table 6: Standard parameter training set results on 50% of the balanced data set. 99

Table 7: Hyperparameter tuned training set results on the imbalanced data set. 100

Table 8: Hyperparameter tuned training set results on the balanced data set. 101

List of Figures

Figure 1: Project timeline. 9

Figure 2: The issue of churn. 12

Figure 3: Model creation flowchart created by author. 36

Figure 4: Histogram of churn instances by Age and Gender. 40

Figure 5: Kendall Rank formula. 42

Figure 6: Kendall’s Rank correlation. 43

Figure 7: Pairplot of variable distribution created by the Author. 45

Figure 8: f-statistic results from ANOVA test. 46

Figure 9: Churn rates of target variable. 49

Figure 10:Target variable class distribution after T-links under sampling. 50

Figure 11: Grid Search vs Random Search. Source: (Bergstra and Bengio 2012) 52

Figure 12: Illustration of the hyperplane space. Source: (Gandhi 2018) 55

Figure 13: The Sigmoid function. Source: (Pant 2019) 56

Figure 14: Factorisation Machines Formula. 57

Figure 15: The Random Forest Algorithm. Source: (Yıldırım 2020) 58

Figure 16: Gradient Boosted Trees learning process. Source: (Yıldırım 2020) 59

Figure 17: Visual representation of the MLP. Source: (Bento 2021) 61

Figure 18: AdaBoost weight calculation. 62

Figure 19: Influence of Decision Stump. 62

Figure 20: Alpha vs Error rate. Source: (McCormick 2013) 62

Figure 21: Extreme Gradient similarity score. 63

Figure 22: Accuracy calculation. 67

Figure 23: Precision Calculation. 68

Figure 24: Recall Calculation. 69

Figure 25: AUC-ROC Calculation. 69

Figure 26: The AUC-ROC evaluation states. Source: (Di Sipio 2021) 70

Figure 27: AdaBoost AUC-ROC plot. 74

Figure 28: Training times of models. 75

Abstract

**Chapter 1: Introduction**

* 1. *Problem Definition*
  2. *Research Objectives*
  3. *Target Variable*

Chapter 2: Methodology

*2.1 Introduction*

*2.2 Data Sourcing and Preparation*

*2.3 Data Limitations*

*2.4 Feature Scaling*

*2.5 Feature Engineering*

*2.6 Feature Selection*

*2.7 Class Imbalance*

*2.8 Hyperparameter Tuning*

*2.9 Model Classifiers*

Decision Tree

Naive Bayes

Support Vector Machine

Logistic Regression

Factorisation Machines Classifier

Random Forest

Gradient Boosted Trees

Multilayer Perceptron Classifier

AdaBoost

Extreme Gradient Boosting

*2.10 Primary Research*

Chapter 3: Results and Discussion

*3.1 Introduction*

*3.2 Evaluation Metrics*

Accuracy

Precision

Recall

F1 score

AUC-ROC

*3.3 Model Results*

*4.4 Conclusion*

References

Appendices

Appendix A

Proof of permission from employment sanctioning the use of data