Configuration Manual

MSc Research Project
MSc AI for Business

Mukarrum Ali Khan
Student ID: x22150269

School of Computing
National College of Ireland

Supervisor: Professor Rejwanul Haque

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Mukarrum Ali Khan.
Student ID:	x22150269.
Programme:	MSc AI for Business **Year:** 2023/2024
Module:	MSc Research Practicum
Supervisor:	Prof Rejwanul Haque
Submission Due Date:	12th August 2024
Project Title:	Transforming the performance of Airline Industry through Sentiment Analysis
Word Count:	797 **Page Count:** 7 pages

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Mukarrum Ali Khan

Date: 12th August 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual

Mukarrum Ali Khan
Student ID: x22150269

1 Introduction:

The tests in this research were run based on the open-source data science AI platform: RapidMiner. This configuration manual provides a comprehensive overview how RapidMiner performs different tests including the data handling procedures, feature engineering, and implementation of machine learning models. It also outlines the system, and its specifications used in this research.

2 System Configuration:

This section explains the hardware and software requirements to implement the machine learning models.

2.1 Hardware Specifications:

Hardware	Build
Machine	Apple MacBook Pro (2020)
Processor	Apple M1 Chip
Ram	8 GB
Storage	256 GB SSD

These hardware specifications are sufficient for handling the dataset used in this research and implementing machine learning algorithms through RapidMiner on them.

2.1 Software Specifications:

macOS Big Slur 11.4 was used in operating data science platform – RapidMiner Studio Educational 10.3.000. RapidMiner is an integrated platform that allows all steps of preparing dataset, training and testing machine learning models, and model deployment without the need of coding or interacting with libraries. RapidMiner in this project was used for data preprocessing, feature engineering, model training and extracting results based on the performance metrics in the research and evaluation.

3 Data Set Preparation:

The dataset used in this research was selected from Kaggle based on a hypothetical company of Airline. The dataset is based on customer reviews, ratings on different services that airline companies provide during the entire journey such as in-flight entertainment, ease of booking, online support, departure or arrival delays. The dataset was pre-processed and prepared for analysis through RapidMiner.

3.1 Dataset Description:

There was total 129,881 customers in the dataset along with features like age, gender, type of travel, class of travel, satisfaction level of customers and ratings of various services based on 0-5 scale. Each service was rated by customers based on their experiences with the airline.

3.2 Dataset Preparation in RapidMiner:

- Data Ingestion: First step before preparation is to ingest the dataset in RapidMiner.
- Data Cleaning: RapidMiner has a built-in tool of handling missing values as the software have capabilities of imputing numerical columns with mean values and categorical values with mode values.
- Encoding: Target encoding, and one-hot coding features are utilized in RapidMiner for categorical values.

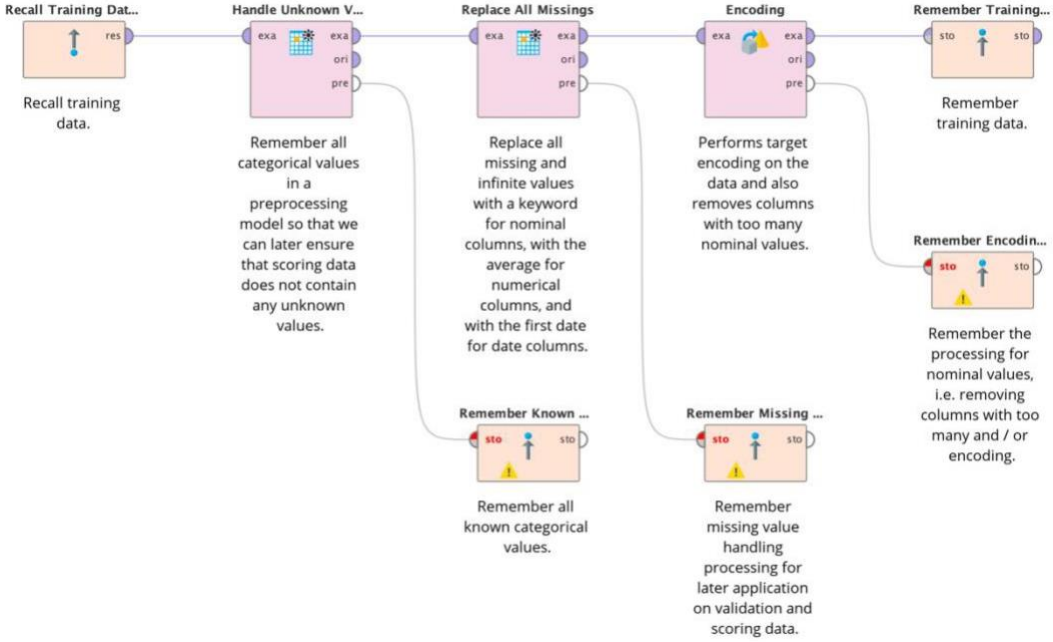

Figure 1: Data Preparation in RapidMiner

4 Implementation:

4.1 Dataset Preprocessing:

After the preparation of dataset for model development, data was pre-processed through RapidMiner to get it ready for further analysis.

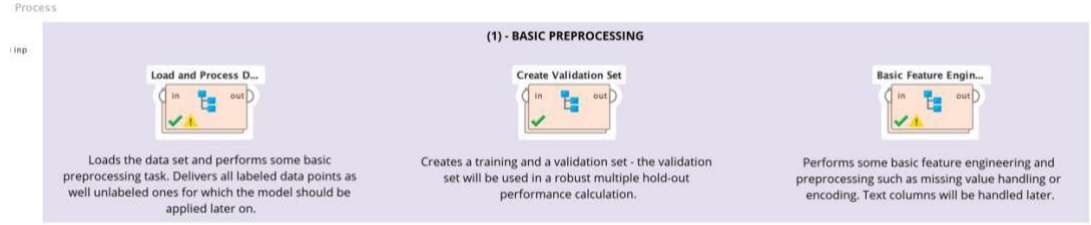

Figure 2: Data Preprocessing in RapidMiner

- Missing Values: Missing values were handled through RapidMiner's imputation operators and ensuring that dataset is ready for analysis.
- Feature Scaling: Numerical Features were standardized to maintain normalcy in data and there are no outliers.
- Text Processing: Customer review texts were processed through different operators of RapidMiner including Tokenization and stemming.

4.2 Feature Extraction:

- Feature Selection: Key features were selected through RapidMiner's automated feature selection tools which are relevant for model training.
- Automatic Feature Engineering: RapidMiner have the capabilities to enhance dataset based on its automatic feature engineering tool to refine the existing features and generate new ones if necessary.

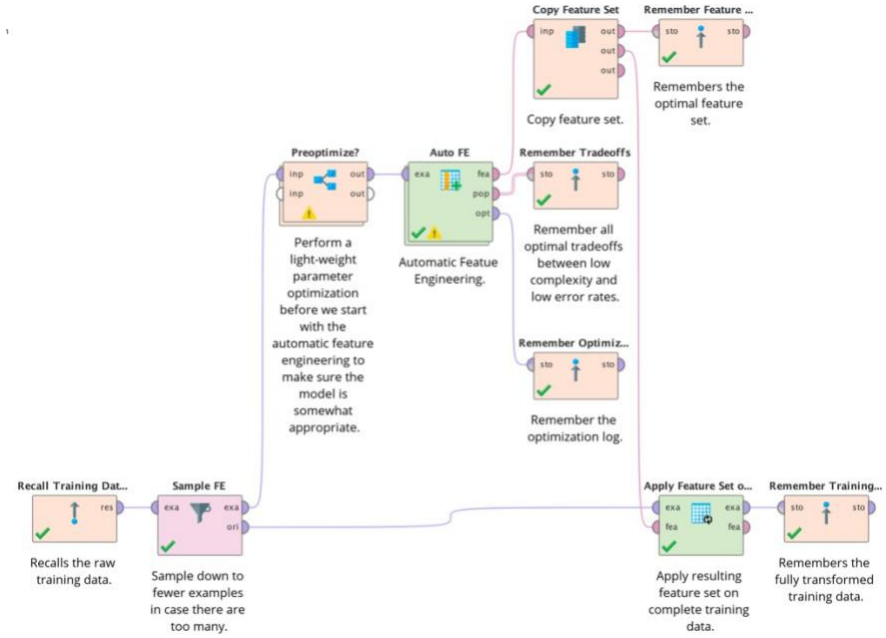

Figure 3: Automatic Feature Engineering of RapidMiner

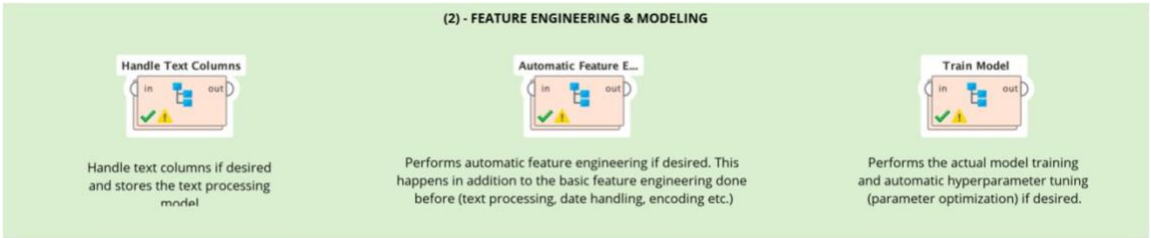

Figure 4: Feature Engineering and Modelling

4.3 Model Training:

Two machine learning models were trained through RapidMiner Auto modelling feature which allows you to incorporate your objective of research and derive results. This research

was based on prediction model as we used RapidMiner in predicting customer satisfaction based on the in-flight services and their ratings in the dataset.

5.1 Random Forest:

Random Forest (RF) is an ensembled machine learning approach which is mainly used for classification and regression tasks. It contains various decision trees and is efficient in handling large datasets (Rane, A. et al., 2018). As this research is based on classification task, RapidMiner grid search operator was used to optimize the parameters and number of trees. This helped in improving the performance of the model. Performance metrics like accuracy, Precision, recall, F1 score, AUC and ROC were evaluated.

5.2 Naïve Bayes:

Naïve Bayes (NB) is a probability classifier algorithm based on Bayes theorem with strong independence between feature assumptions. Naïve Bayes scalability is one of the key factors that makes it highly credible for classification tasks. It's a very famous method in text categorization because of its efficiency and simplicity (Melville et al., 2009). It's another classification algorithm used in this research as a baseline model for comparison. Similar metrics were used in evaluating both the models.

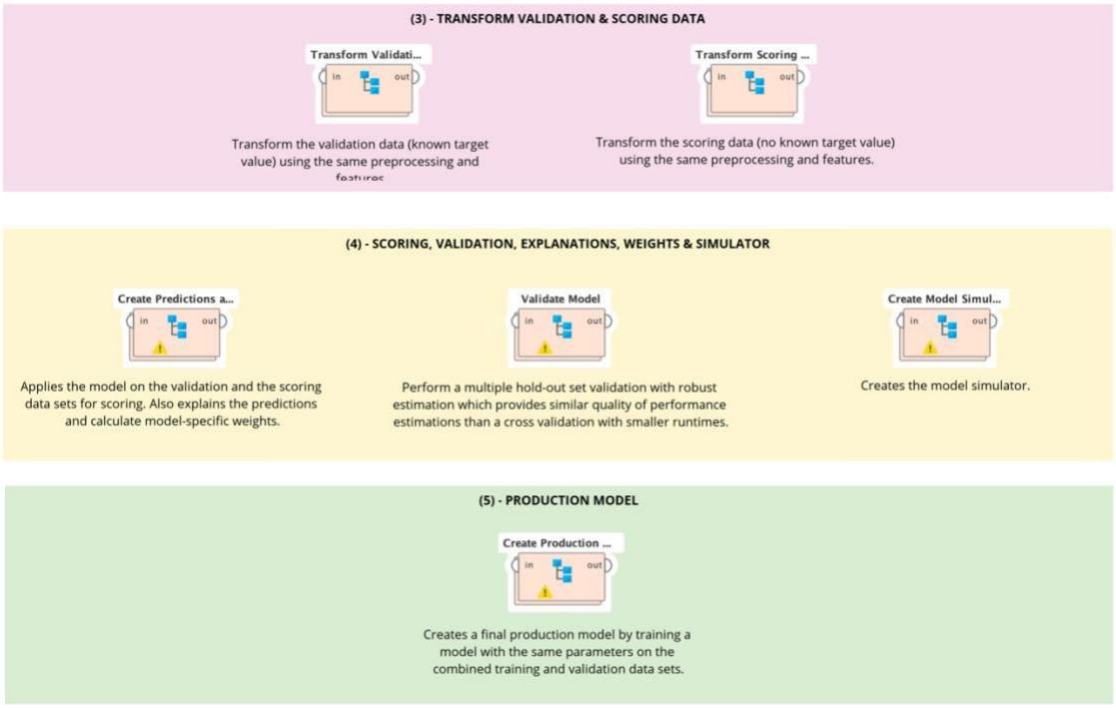

Figure 5: Steps of Model Training and Production

5 Evaluation:

Random Forest and Naïve Bayes were deployed in this research to run sentiment analysis and gather meaningful insights that airline industries could work on, transform their business performance and enhance the experience for customers. RF outperformed NB as it turned out to be a better machine learning tool in this research. Figure 6 explains the results derived from implementing both models through RapidMiner:

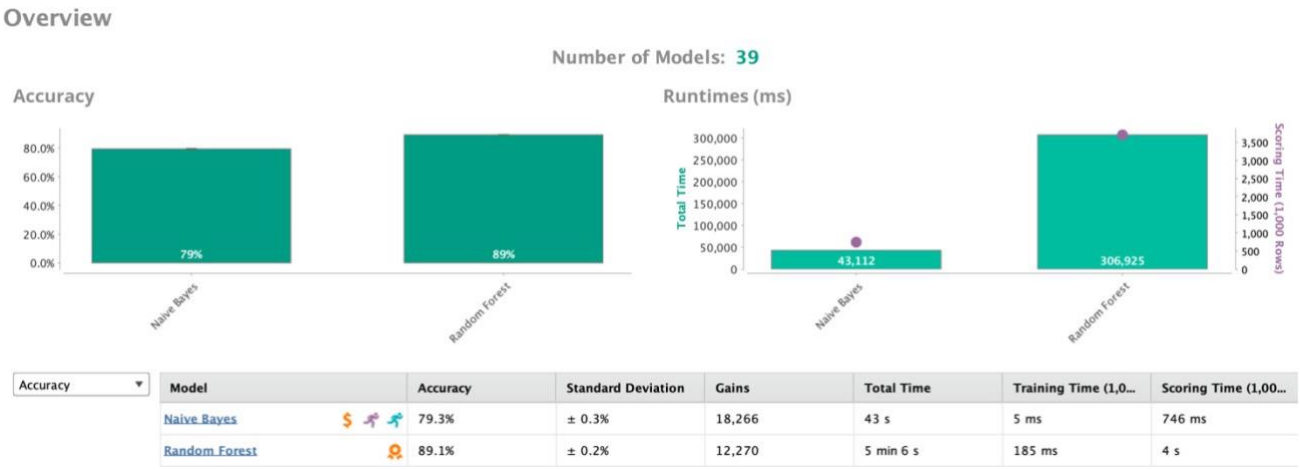

Figure 6: Results

References

Melville, Prem, Wojciech Gryc, and Richard D. Lawrence. "Sentiment analysis of blogs by combining lexical knowledge with text classification." Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2009.

Rane, A. and Kumar, A., 2018, July. Sentiment classification system of twitter data for US airline service analysis. In 2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC) (Vol. 1, pp. 769-773). IEEE.