Eduvest – Journal of Universal Studies
Volume 5 Number 1, January, 2025
p- ISSN 2775-3735- e-ISSN 2775-3727

CHURN PREDICTION ANALYSIS OF CUSTOMER FERRY OPERATOR "BATAMFAST" USING MACHINE LEARNING WITH SUPERVISED CLASSIFICATION MODEL

Ryan Tri Pamungkas[1]*, Nenden Siti Fatonah[2], Gerry Firmansyah[3], Budi Tjahjono[4]
Universitas Esa Unggul, Jakarta, Indonesia [1,2,3,4]
Email: ryantripamungkas@gmail.com

ABSTRACT

BatamFast is the first ferry operator in Batam and has been serving international routes to Singapore and Malaysia since 1985. Until 2010, BatamFast was the only ferry operator in Batam. However, since 2011, competitors such as Sindo Ferry, Horizon Ferry, and Majestic have emerged, increasing competition to four ferry operators in Batam by 2024. This study aims to measure the churn rate of BatamFast customers and identify the factors causing it using machine learning prediction models such as Random Forest, XGBoost, and Gradient Boosting. In addition to churn prediction, feature importance analysis was conducted to determine the significant features influencing customer decisions. The results indicate that XGBoost is the best model compared to Random Forest and Gradient Boosting. Key factors for churn are customer category, payment method, and booking mode. These findings are expected to help BatamFast reduce churn, improve customer satisfaction, and strengthen its competitive position in the international ferry market.

KEYWORDS *Churn Prediction, Batamfast, machine learning, feature importance, Random Forest, XGBoost, Gradient Boosting*

INTRODUCTION

BatamFast is an international ferry operator serving the Batam - Singapore route, established in 1985, and was the first ferry operator in Batam with the Batam - Singapore route and became the only ferry operator with the route from 1985 to 2010 (Osman & Ghaffari, 2021). Over time, several other international ferry operators are present to serve the same route as BatamFast, namely Sindo Ferry in 2011, Horizon Ferry and Majestic ferry in 2014 (Mohammad et al., 2019). Until 2024, there are four International Ferry Operators on the Batam - Singapore route (Liu et al., 2021).

How to cite:	Ryan Tri Pamungkas, et al. (2025). Churn Prediction Analysis Of Customer Ferry Operator "Batamfast" Using Machine Learn-Ing With Supervised Classification Model. *Journal Eduvest.* 5(1), 1144-1154
E-ISSN:	2775-3727

With the many choices of ferry operators in Batam, the possibility of customers switching to other ferry operators is a concern for BatamFast (Prabadevi et al., 2023). This is important to retain customers who are already using BatamFast services and attract new customers (Ahmad et al., 2019).

Customer churn refers to a company's customer attrition rate, which is one of the most important metrics for growing businesses to evaluate (Kim & Lee, 2022). A considerable reason for customer churn activity is dissatisfaction with existing services (Nurhidayat & Anggraini, 2023). Based on these problems, this study aims to identify the most significant factors or features by conducting a feature importance factor analysis, with the analysis being able to find out the factors that affect the customer's decision to switch to another ferry operator (Shrestha & Shakya, 2022). By understanding these factors, it is hoped that companies can take strategic steps to reduce churn and increase customer satisfaction (Loria & Marconi, 2021).

And also evaluate the method used in this study, namely machine learning and focus on several supervised classification models, namely Random Forest, XGBoost, and Gradient Boosting Model, From the evaluation, it is hoped that the classification model that has the best performance can be determined, and this research is expected to provide insight into the factors that affect churn and help the company in designing an effective strategy to retain customers (Dhangar & Anand, 2021). The objectives of this study are 1 (De Caigny et al., 2020). Evaluate the performance of the Random Forest, XGBoost, and Gradient Boosting Model classification models in predicting customer churn activity (Jain et al., 2020; Raeisi & Sajedi, 2020).

RESEARCH METHOD

This research was carried out with a quantitative approach and this research was carried out from March to August 2024. The object of the research is the BatamFast ferry operator company. It is the main source of data to obtain information about the problems that researchers need to discuss in this study.

Batam Fast Ferry Pte Ltd was established in 1985. Initially, the company had purchased 2 high-speed passenger ferries, namely Bintan 2 and Bintan 3 with a capacity of 60 passengers and 70 passengers respectively as the beginning of its business operations. transporting passengers from Singapore to Batam Island and vice versa (Bhuse et al., 2020).

Data is collected from various internal sources of the company such as annual reports and internal databases. The collected data will then become materials to be processed and analyzed according to research needs. In the research method, the author takes reference by following the stages of the Cross-Industry Standard Process for Data Mining (CRISP-DM) model (Agarwal et al., 2022). The stages of CRISP-DM are Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment.

RESULT AND DISCUSSION

Business Understanding

In this study, the business understanding stage is an important first step to understand the context of the problems faced by the BatamFast ferry operator company. BatamFast faces challenges in maintaining the number of passengers and reducing churn (customers who stop using services). Therefore, an in-depth understanding of these business issues is necessary to identify the factors influencing churn and devise effective strategies to improve customer retention.

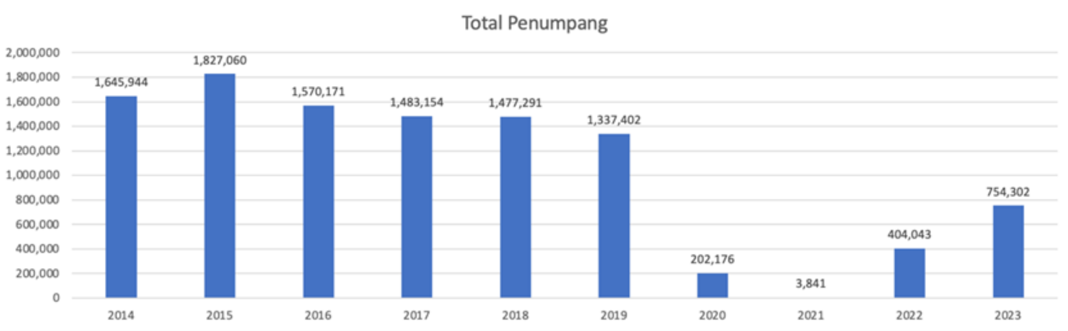


Figure 1 Total Passengers

Based on the total passenger data in Figure 1, the number of passengers experienced an upward trend in the years 2014 to 2015. However, there were several periods where the number of passengers declined in 2016, 2017, 2018, 2019, 2020, 2021.

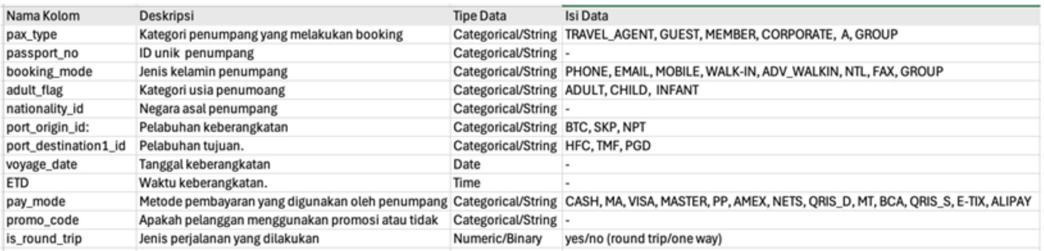
The decline in the number of passengers in 2016 to 2019 is likely due to several factors that can occur, such as economic conditions and an increase in the number of ferry operators with the same route as BatamFast which adds to competition in the ferry operator industry.

And the decline in the number of passengers in 2020 to 2021 is most likely due to the COVID-19 pandemic.

Data Understading

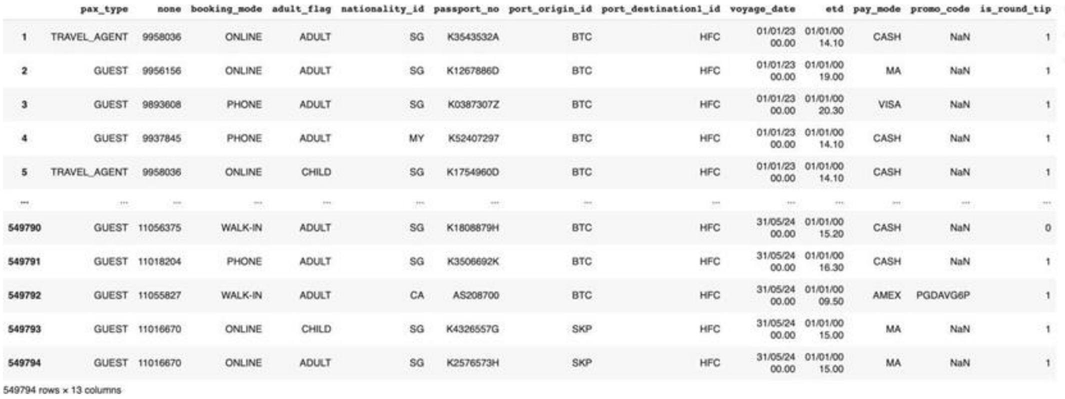
In this study, the Data Understanding stage is a crucial step to ensure that the available data is relevant and of good quality to support analysis and modeling. The data collected includes a variety of important information related to customers and their transactions with BatamFast.

Dataset Description



Nama Kolom	Deskripsi	Tipe Data	Isi Data
pax_type	Kategori penumpang yang melakukan booking	Categorical/String	TRAVEL_AGENT, GUEST, MEMBER, CORPORATE, A, GROUP
passport_no	ID unik penumpang	Categorical/String	-
booking_mode	Jenis kelamin penumpang	Categorical/String	PHONE, EMAIL, MOBILE, WALK-IN, ADV_WALKIN, NTL, FAX, GROUP
adult_flag	Kategori usia penumoang	Categorical/String	ADULT, CHILD, INFANT
nationality_id	Negara asal penumpang	Categorical/String	-
port_origin_id:	Pelabuhan keberangkatan	Categorical/String	BTC, SKP, NPT
port_destination1_id	Pelabuhan tujuan.	Categorical/String	HFC, TMF, PGD
voyage_date	Tanggal keberangkatan	Date	-
ETD	Waktu keberangkatan.	Time	-
pay_mode	Metode pembayaran yang digunakan oleh penumpang	Categorical/String	CASH, MA, VISA, MASTER, PP, AMEX, NETS, QRIS_D, MT, BCA, QRIS_S, E-TIX, ALIPAY
promo_code	Apakah pelanggan menggunakan promosi atau tidak	Categorical/String	-
is_round_trip	Jenis perjalanan yang dilakukan	Numeric/Binary	yes/no (round trip/one way)

The data used in this study was obtained from various internal sources of the company, including annual reports and internal databases. The raw dataset used has 549515 rows and 12 columns. This information provides an initial idea of the size and complexity of the data to be analyzed using Python.

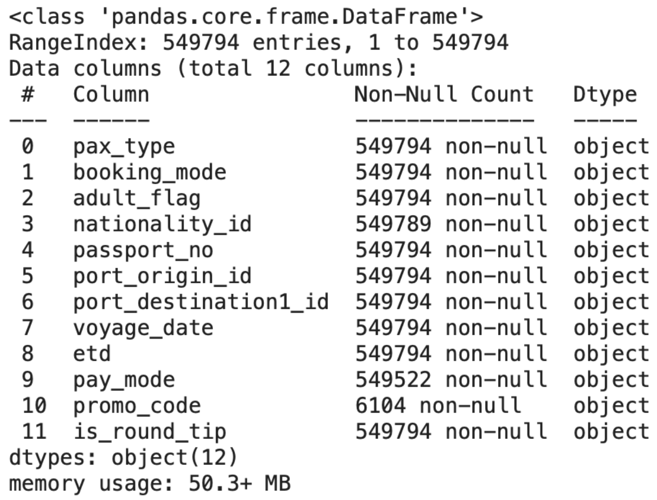


	pax_type	none	booking_mode	adult_flag	nationality_id	passport_no	port_origin_id	port_destination1_id	voyage_date	etd	pay_mode	promo_code	is_round_tip
1	TRAVEL_AGENT	9958036	ONLINE	ADULT	SG	K3543532A	BTC	HFC	01/01/23 00.00	01/01/00 14.10	CASH	NaN	1
2	GUEST	9956156	ONLINE	ADULT	SG	K1267886D	BTC	HFC	01/01/23 00.00	01/01/00 19.00	MA	NaN	1
3	GUEST	9893608	PHONE	ADULT	SG	K0387307Z	BTC	HFC	01/01/23 00.00	01/01/00 20.30	VISA	NaN	1
4	GUEST	9937845	PHONE	ADULT	MY	K52407297	BTC	HFC	01/01/23 00.00	01/01/00 14.10	CASH	NaN	1
5	TRAVEL_AGENT	9958036	ONLINE	CHILD	SG	K1754960D	BTC	HFC	01/01/23 00.00	01/01/00 14.10	CASH	NaN	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...
549790	GUEST	11056375	WALK-IN	ADULT	SG	K1808879H	BTC	HFC	31/05/24 00.00	01/01/00 15.20	CASH	NaN	0
549791	GUEST	11018204	PHONE	ADULT	SG	K3506692K	BTC	HFC	31/05/24 00.00	01/01/00 16.30	CASH	NaN	1
549792	GUEST	11055827	WALK-IN	ADULT	CA	AS208700	BTC	HFC	31/05/24 00.00	01/01/00 09.50	AMEX	PGDAVG6P	1
549793	GUEST	11016670	ONLINE	CHILD	SG	K4326557G	SKP	HFC	31/05/24 00.00	01/01/00 15.00	MA	NaN	1
549794	GUEST	11016670	ONLINE	ADULT	SG	K2576573H	SKP	HFC	31/05/24 00.00	01/01/00 15.00	MA	NaN	1

549794 rows × 13 columns

Figure 2. Initial Data Sample

In Figure 2. there are 13 sample rows with various characteristics in each column, but it can be seen that there are still columns that have NaN values, namely Missing Values or Not a Number, this will be done to clean the data so that the raw data becomes better quality data.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 549794 entries, 1 to 549794
Data columns (total 12 columns):
 #   Column                Non-Null Count   Dtype 
---  ------                --------------   ----- 
 0   pax_type              549794 non-null  object
 1   booking_mode          549794 non-null  object
 2   adult_flag            549794 non-null  object
 3   nationality_id        549789 non-null  object
 4   passport_no           549794 non-null  object
 5   port_origin_id        549794 non-null  object
 6   port_destination1_id  549794 non-null  object
 7   voyage_date           549794 non-null  object
 8   etd                   549794 non-null  object
 9   pay_mode              549522 non-null  object
 10  promo_code            6104 non-null    object
 11  is_round_tip          549794 non-null  object
dtypes: object(12)
memory usage: 50.3+ MB
```

Figure 3. Data Set Information

In Figure 4.3, it can be seen that the dataset information still has many rows that have missing values or are blank. This can make the data impossible to process in Machine Learning. The same goes for the type or data type of each row. All data types on each row are still of type object data.

This makes the data inconsistent in the analysis process, because the type of data that should be numerical cannot be processed in a mathematical formula. This will be done by converting data types at the data pre-processing stage.

Data Preparation

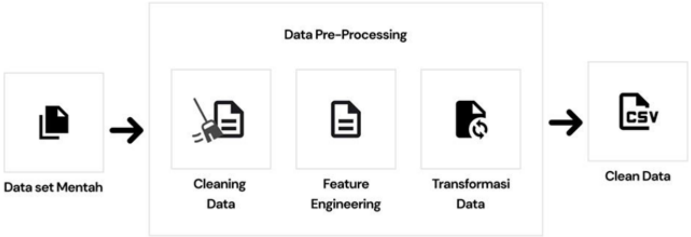
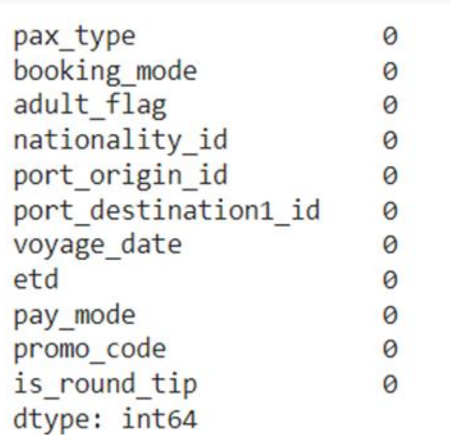


Figure 4. Data Preparation

Handling Missing and Duplicate Values

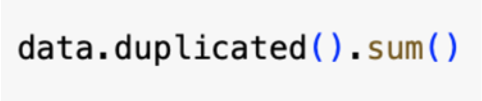
1. Handling missing values The missing values will be filled in using the appropriate method, there are 3 columns that have missing values, namely promo_code, pay_mode, and nationality_id.
2. Calculate the amount of value lost after handling
3. Reinstating the value of nothing is lost



```
pax_type               0
booking_mode           0
adult_flag             0
nationality_id         0
port_origin_id         0
port_destination1_id   0
voyage_date            0
etd                    0
pay_mode               0
promo_code             0
is_round_tip           0
dtype: int64
```

Figure 5. Stages of Feature Engineering

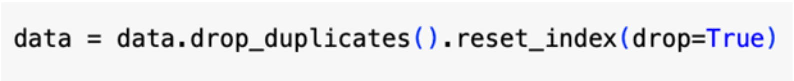
4. Count the number of duplicate rows



```
data.duplicated().sum()
```

5. Handling 2 duplicate lines found, by deleting and leaving one of the rows

By using:



```
data = data.drop_duplicates().reset_index(drop=True)
```

Then the rows will be no more duplicates.

Feature Engineering

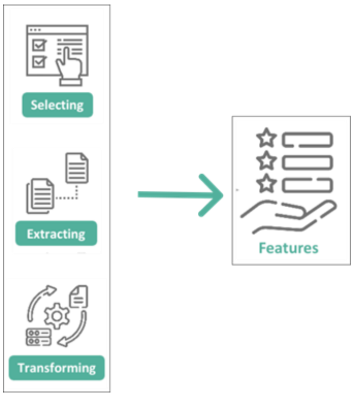


Figure 6. Feature Engineering Process

In this feature engineering process, the goal is to create and process features from raw data to help machine learning models work better. This feature becomes the information that the model uses to make predictions.

Replacing Values and Encoding

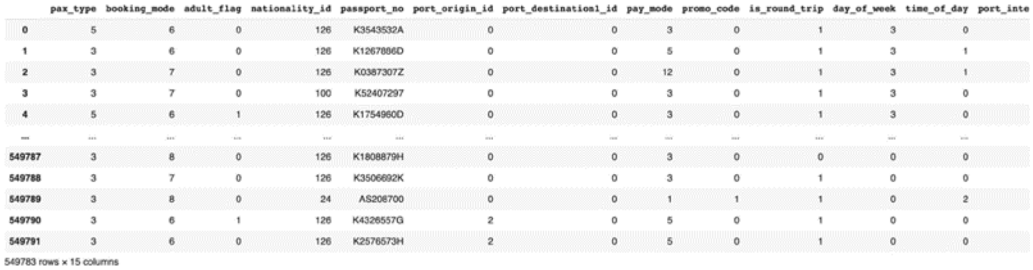
1. Change promo_code columns

   This changes all values other than 'No Promo Code' to 1, and the value of 'No Promo Code' to 0, which indicates whether the promo code is used (1) or not used (0). Changing the promo_code data type to int, After changing the value, the promo_code column data type is changed to an integer (int). This ensures that all values in the column are integers (0 or 1).
2. Data encoding

   Removing the voyage_date and etd columns from the DataFrame as they may no longer be needed for further analysis or machine learning models, Create an instance of LabelEncoder from scikit-learn. LabelEncoder is used to convert category labels to numbers, Using LabelEncoder to convert values. After converting, the data type of the columns is changed to an integer (int).

Display of data after cleaning

After going through the cleaning stage, here is a view of the data that has been cleaned:



	pax_type	booking_mode	adult_flag	nationality_id	passport_no	port_origin_id	port_destination1_id	pay_mode	promo_code	is_round_trip	day_of_week	time_of_day	port_inte
0	5	6	0	126	K3543532A	0	0	3	0	1	3	0	
1	3	6	0	126	K1267886D	0	0	5	0	1	3	1	
2	3	7	0	126	K0387307Z	0	0	12	0	1	3	1	
3	3	7	0	100	K52407297	0	0	3	0	1	3	0	
4	5	6	1	126	K1754960D	0	0	3	0	1	3	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	
549787	3	8	0	126	K1808879H	0	0	3	0	0	0	0	
549788	3	7	0	126	K3506692K	0	0	3	0	1	0	0	
549789	3	8	0	24	AS206700	0	0	1	1	1	0	2	
549790	3	6	1	126	K4326557G	2	0	5	0	1	0	0	
549791	3	6	0	126	K2576573H	2	0	5	0	1	0	0	

549783 rows × 15 columns

Figure 7. Clean data

Exploratory Data Analysis

The following visualization of churn distribution in figure 4.8 shows the number of customers Not Churn (0): 488,330 Number of churn customers (1): 61.45.

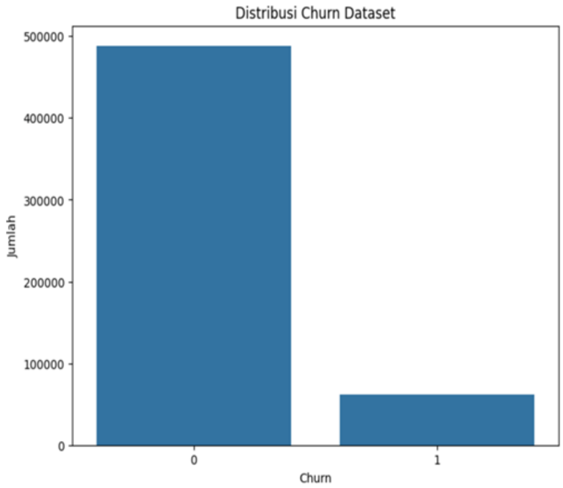


Figure 8. Churn Distribution

Modeling

In this study, churn prediction modeling was carried out using three different machine learning algorithms, namely:

1. Random Forest
2. XGBosst
3. Gradient Boosting

And using hyperparameters with Manual Tuning on each model, Hyperparameters are used to optimize the performance of the model used. This modeling process is to build and train models to understand or predict patterns in data. Each model is trained and evaluated using pre-cleaned and prepared data.

Results of Evaluation Analysis

Table 1. Model Evaluation Comparison

Type	Precision (Class 0)	Recall (Class 0)	F1-score (Class 0)	Precision (Class 1)	Recall (Class 1)	F1-score (Class 1)	Accuracy
Random Forest	0.91	0.99	0.94	0.63	0.18	0.28	0.90
XGBoost	0.91	0.98	0.94	0.60	0.27	0.37	0.90
Gradient Boosting	0.91	0.72	0.81	0.16	0.43	0.24	0.69

From table 1, it can be seen that the comparison of the evaluation of each model is as follows:

1. Random Forest

Advantages

a. Precision and recall for Class 0 is very high
b. Precision and recall for Class 1 are low, with F1-score also low.

Deficiency

a. Precision and recall for Class 1 are low, with F1-score also low.

2. XGBoost

Advantages

a. Precision and recall for Class 0 is very high
b. Accuracy is equivalent to Random Forest.

Deficiency

a. Precision and recall for Class 1 are low, with F1-score also low.
b. Precision and recall for Class 1 are still low, but still slightly better than other models

3. Gradient Boosting

Advantages

a. Has high precision.

Deficiency

a. Precision and recall for Class 1 are very low.
b. Accuracy is very low compared to Random Forest and XGBoost.

Table 2. Comparison of Confusion Matrix Models

Type	MR	FP	FN	TP
Random Forest	120,486	1,684	12,467	2,809
XGBoost	119,39	2,780	11,149	4,127
Gradient Boosting	88,386	33,784	8,69	6,586

From the table above, we can see the confusion matrix of each model as follows:

1. Random Forest: Good at detecting non-churn (very little FP) but not good at detecting churn (a lot of FN).
2. XGBoost: Balanced in detecting churn and non-churn. The FP is slightly higher than Random Forest, but it has a higher TP, so it is better at detecting churn.
3. Gradient Boosting: Detects churn very well (high TP), but too many errors in detecting non-churn (very high FP).

Feature Importance & Partial Dependence Plots Results

Permutation feature importance measures the importance of each feature by randomizing its value and observing the deterioration of model performance. The following is the implementation in this study.

1. Starting with the machine learning models that have been trained, namely Random Forest, XGBosst, Gradient Boosting using training data. The model has learned the relationship between features and targets based on training data.
2. Measure model performance on original test data by calculating evaluation metrics such as accuracy, F1-score
3. Each feature that wants to be analyzed will be randomized to the value of the feature on the test data, by changing the value of the feature to random, while the value of other features remains at the original value.
4. Evaluating Performance After Permutation.
   Using modified test data (with shuffled feature values), the model is run and calculates the model's performance on that data.
5. Calculating Average Performance Drop
   Repeat the permutation process several times to get a stable estimate of the performance degradation.

6. Feature Rating

   The feature that causes the greatest drop in performance when the values are randomized is considered the most important.

   Features that cause a small drop in performance are considered less important.

Here are some of the results of feature importance that have been analyzed from 3 models, namely XGBoost, Random Fores, and Gradient Boosting.

1. XGBoost

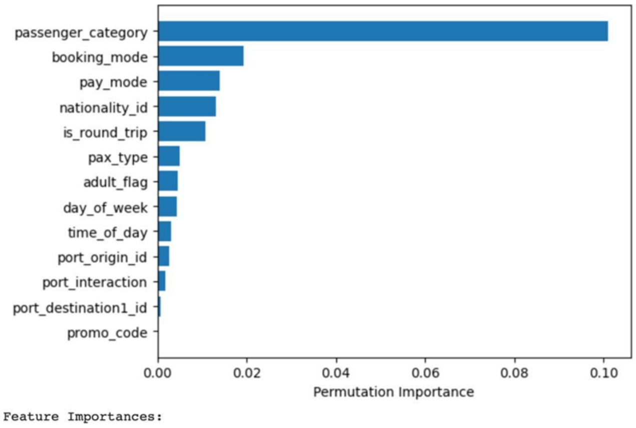


Figure 9. Feature Importance XGBoost

So you can get the 3 highest features that allow churn to occur, namely passenger_category, booking_mode and pay_mode

2. Random Forest

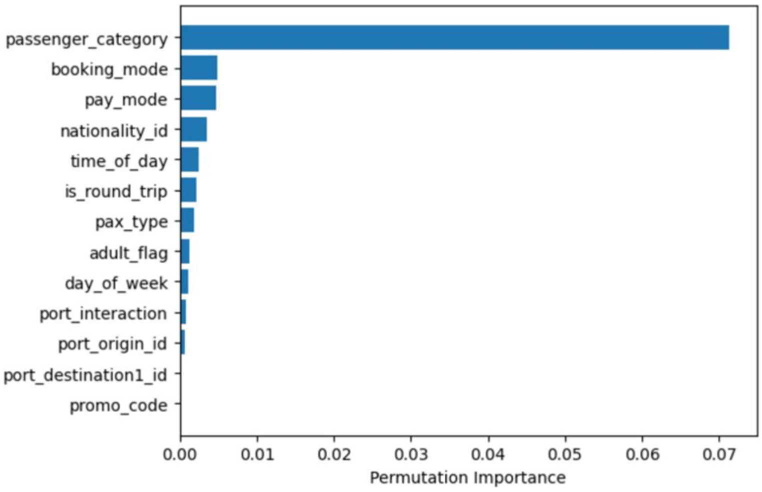


Figure 10. Feature Importance Random Forest

So you get the 3 highest features that allow churn to occur, namely passenger_category, booking_mode and pay_mode the same as XGBoost but XGBoost has a higher value.

3. Gradient Boosting

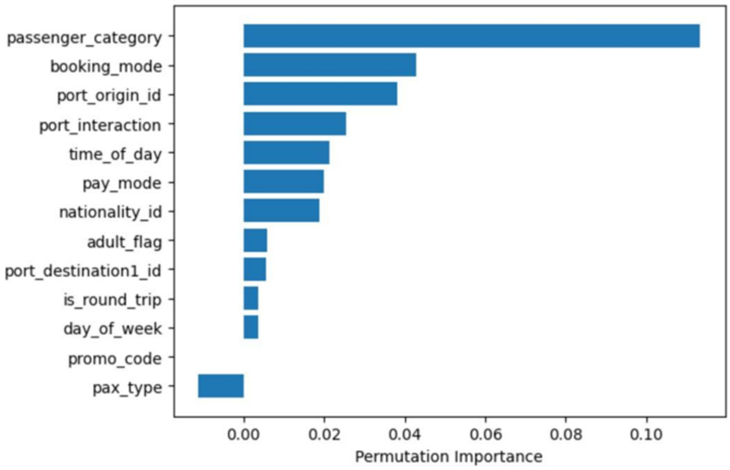


Figure 11. Feature Importance Gradient Boosting

So the 3 highest features that allow churn to occur are passenger_category, booking_mode and port_origin_id, in The feature importance of this model has a fairly high value, but for this gradient boosting model, the accuracy is very low.

CONCLUSION

Based on the research that has been conducted on the analysis of BatamFast customer chrun predictions using machine learning with a supervised clasification model, here are some conclusions that can be drawn, namely XGBoost is the most effective model in predicting churn compared to Random Forest and Gradient Boosting. This model shows the best results in terms of precision, recall, and f1-score. The accuracy of this model is 0.90, with the highest f1-score for the churn category at 0.37. This indicates that XGBoost is able to better capture important patterns that indicate churn.

Based on the results of feature importance and Partial Dependence Plots, it can be known what factors affect churn, namely, passenger_category, booking_mode, and pay_mode have the greatest contribution to the likelihood of churn. The passenger_category feature had the highest importance value of 0.1010, with Returning Passenger showing the most significant impact on churn, with a PDP value of 0.3362. booking_mode and pay_mode also contributed significantly with important values of 0.0193 and 0.0139, respectively. In PDP, the NTL (Not To Land) for booking_mode shows a value of 0.1858, while the QRIS_S for pay_mode shows a PDP value of 0.2117, which indicates the largest contribution in possible churn among the analyzed payment methods.

REFERENCES

Agarwal, V., Taware, S., Yadav, S. A., Gangodkar, D., Rao, A. L. N., & Srivastav, V. K. (2022). Customer-Churn Prediction Using Machine Learning. 2022 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS), 893–899.

Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. Journal of Big Data, 6(1), 1–24.

Bhuse, P., Gandhi, A., Meswani, P., Muni, R., & Katre, N. (2020). Machine learning based telecom-customer churn prediction. 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), 1297–1301.

De Caigny, A., Coussement, K., De Bock, K. W., & Lessmann, S. (2020). Incorporating textual information in customer churn prediction models based on a convolutional neural network. International Journal of Forecasting, 36(4), 1563–1578.

Dhangar, K., & Anand, P. (2021). A Review on Customer Churn Prediction Using Machine Learning Approach. International Journal of Innovations in Engineering Research and Technology, 8(05), 193–201.

Jain, H., Khunteta, A., & Srivastava, S. (2020). Churn prediction in telecommunication using logistic regression and logit boost. Procedia Computer Science, 167, 101–112.

Kim, S., & Lee, H. (2022). Customer churn prediction in influencer commerce: An application of decision trees. Procedia Computer Science, 199, 1332–1339.

Liu, K., Hu, X., Zhou, H., Tong, L., Widanage, W. D., & Marco, J. (2021). Feature analyses and modeling of lithium-ion battery manufacturing based on random forest classification. IEEE/ASME Transactions on Mechatronics, 26(6), 2944–2955.

Loria, E., & Marconi, A. (2021). Exploiting limited players' behavioral data to predict churn in gamification. Electronic Commerce Research and Applications, 47, 101057.

Mohammad, N. I., Ismail, S. A., Kama, M. N., Yusop, O. M., & Azmi, A. (2019). Customer churn prediction in telecommunication industry using machine learning classifiers. Proceedings of the 3rd International Conference on Vision, Image and Signal Processing, 1–7.

Nurhidayat, M. M. S., & Anggraini, D. (2023). Analysis and Classification of Customer Churn Using Machine Learning Models. Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi), 7(6), 1253–1259.

Osman, Y., & Ghaffari, B. (2021). Customer churn prediction using machine learning: A study in the B2B subscription based service context.

Prabadevi, B., Shalini, R., & Kavitha, B. R. (2023). Customer churning analysis using machine learning algorithms. International Journal of Intelligent Networks, 4, 145–154.

Raeisi, S., & Sajedi, H. (2020). E-commerce customer churn prediction by gradient boosted trees. 2020 10th International Conference on Computer and Knowledge Engineering (ICCKE), 55–59.

Shrestha, S. M., & Shakya, A. (2022). A customer churn prediction model using XGBoost for the telecommunication industry in Nepal. Procedia Computer Science, 215, 652–661.