DATA SCIENCE INDONESIA

OFFERING AN END-TO-END SOLUTION : DATA-DRIVEN STRATEGIES FOR A COMPREHENSIVE APPROACH TO CUSTOMER CHURN

DATA GARIS KERAS (DGK)

DATA GARIS KERAS

NAME : Putri Awalia Shabrina
MAJOR : Physics
UNIVERSITY : Gadjah Mada University

NAME : Zulfikar Irham
MAJOR : Physics
UNIVERSITY : Gadjah Mada University

NAME : Shahnaz Izzati Frishila
MAJOR : Electronic and Instrumentation
UNIVERSITY : Gadjah Mada University

Introduction

Telecommunications is a rapidly growing company in terms of technology, level of competition, number of operators, services, and so on. Due to the fierce competition, the factors that increase the tendency of customers to abandon a product need to be studied comprehensively.

The Pareto Principle in business refers to 80 percent of business profits coming from 20 percent of valuable customers. To retain valuable customers, it is important for companies to understand the customer behavior based on the given data.

DATA & METHOLOGY

DATA

- Data has 7043 customer behavior on telco company
- Data has 7043 rows and 16 features

DATA UNDERSTANDING

Numerical Features

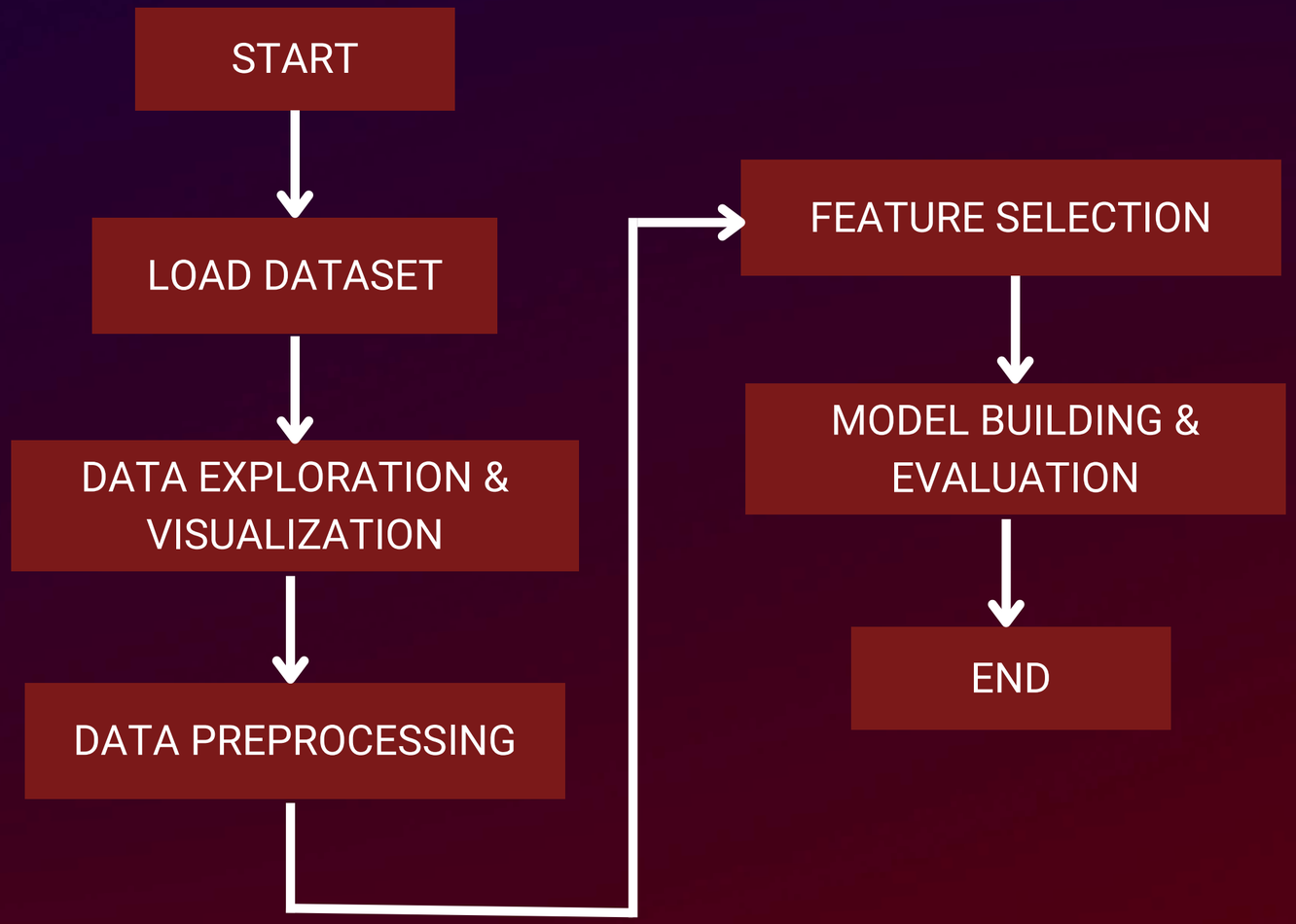
- **Customer ID** (A unique ID for each customer)
- **Tenure Months** (How long customers stay in the company)
- **Monthly Purchase** (Total customer’s monthly spent for all services with the unit of **thousands of IDR**)
- **Longitude** (Customer’s residence - Longitude)
- **Latitude** (Customer’s residence - Latitude)
- **CLTV** (Customer Lifetime Value with the unit of **thousands of IDR** - Calculated using company's formulas)

DATA UNDERSTANDING

Categorical Features

- **Device Class** (Device classification)
- **Games Product** (Whether the customer uses the internet service for games product)
- **Music Product** (Whether the customer uses the internet service for music product)
- **Education Product** (Whether the customer uses the internet service for education product)
- **Call Center** (Whether the customer uses the call center service)
- **Video Product** (Whether the customer uses video product service)
- **Use MyApp** (Whether the customer uses MyApp service)
- **Payment Method** (The method used for paying the bill)
- **Churn Label** (Whether the customer left the company in this quarter)

Methology



Loading the Dataset: using pandas' `read_excel` function and stores it in a DataFrame called `df`.

Data Exploration and Visualization: data exploration and visualization tasks such as checking for unique values

Data Preprocessing:

- Label encoding is performed on the target variable 'Churn Label' and categorical features using `LabelEncoder`
- Feature scaling is done using `MinMaxScaler` from sklearn.

Feature Selection: Chi-squared test and ANOVA test are used for feature selection to identify the most important features for the prediction model.

Model Building and Evaluation: code splits the data, builds multiple classification, then Model performance is evaluated

Data Exploratory & Visualization

- Looking for the shape data obtained is 7043 rows and 16 columns.
- Check Null values and duplicated values. Not found null values and duplicated values.
- determine the target variable, namely Churn Label
- Dissecting the data for each feature determines whether this data is categorical or numeric. If it is numerical, a histogram visualization of the target is carried out. If the category is done, data visualization is carried out in the form of a chart against the target.

Data Preprocessing

Location	Device Class	Games Product	Music Product	Education Product	Call Center	Video Product	Use MyApp	Payment Method
1	2	2	2	0	0	0	0	2
1	0	0	0	0	0	0	0	3
1	0	0	0	2	0	2	2	3
1	0	0	0	2	1	2	2	3
1	0	0	2	2	0	2	2	1

Tenure Months	CLTV (Predicted Thou. IDR)	Monthly Purchase (Thou. IDR)
0.027778	0.274850	0.354229
0.027778	0.155215	0.521891
0.111111	0.749166	0.809950
0.388889	0.667111	0.861194
0.680556	0.742050	0.850249

- Label encoding data using LabelEncoder in sklearn. Change the categorical data to numerical data with data type integer (0,1,2,...).
- Feature scaling using MinMaxScaler in sklearn. Scale numerical features in range from 0 to 1.

Feature Selection

Chi-square

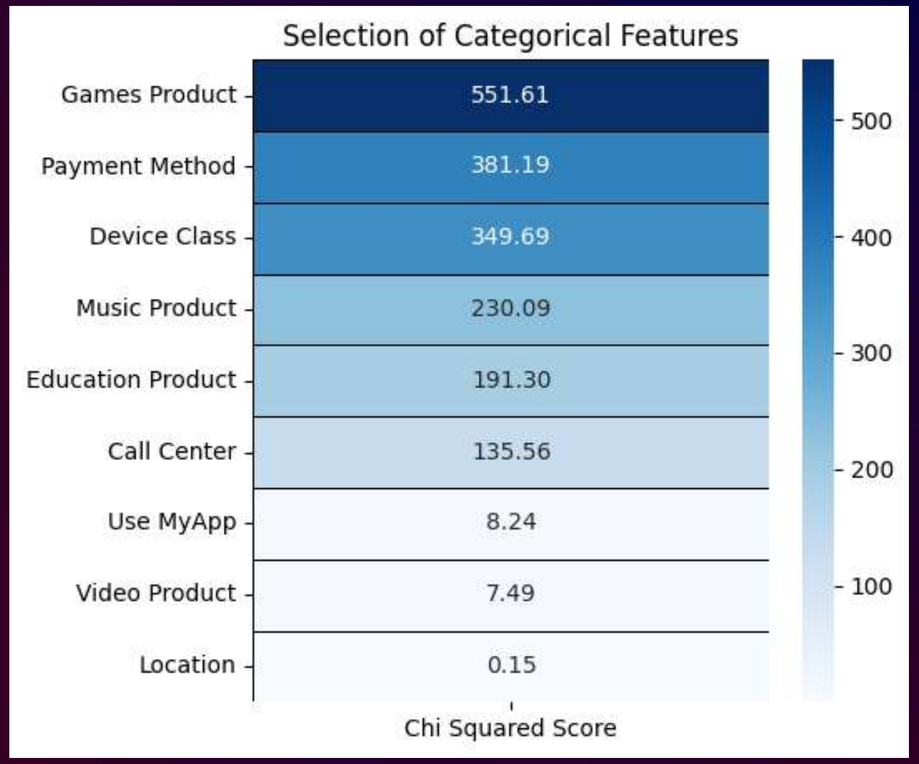
To determine the relationship between categorical features and target feature

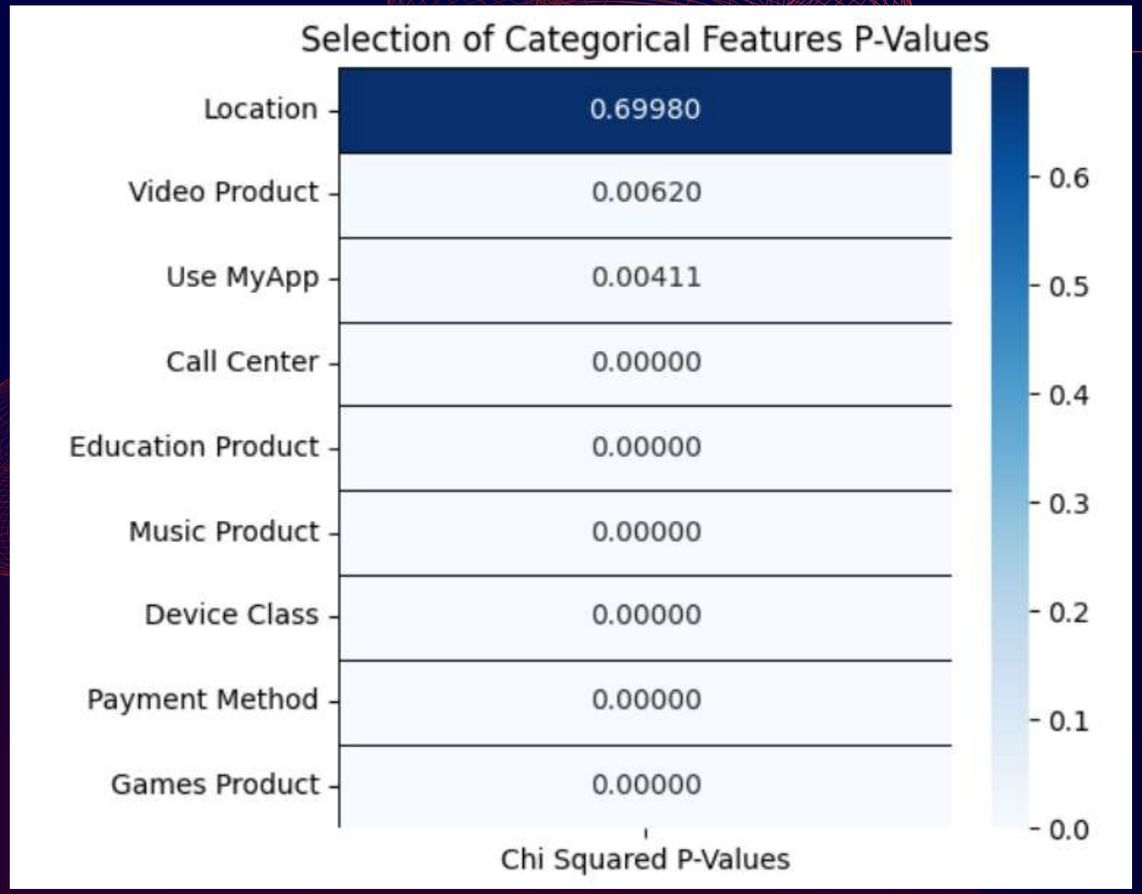
Define hypothesis

- Null Hypothesis (H0): There is no significant relationship between the categorical features and the target feature Churn Label.
- Alternate Hypothesis (H1): is that there is a significant relationship between the categorical features and the target feature Churn Label.

Feature Selection

Chi-square





If p-values > 0.05 we can reject the null hypothesis. Feature location is not related to Churn. So we don't use feature Location for modelling

Feature Selection

ANOVA (Analysis of Variance

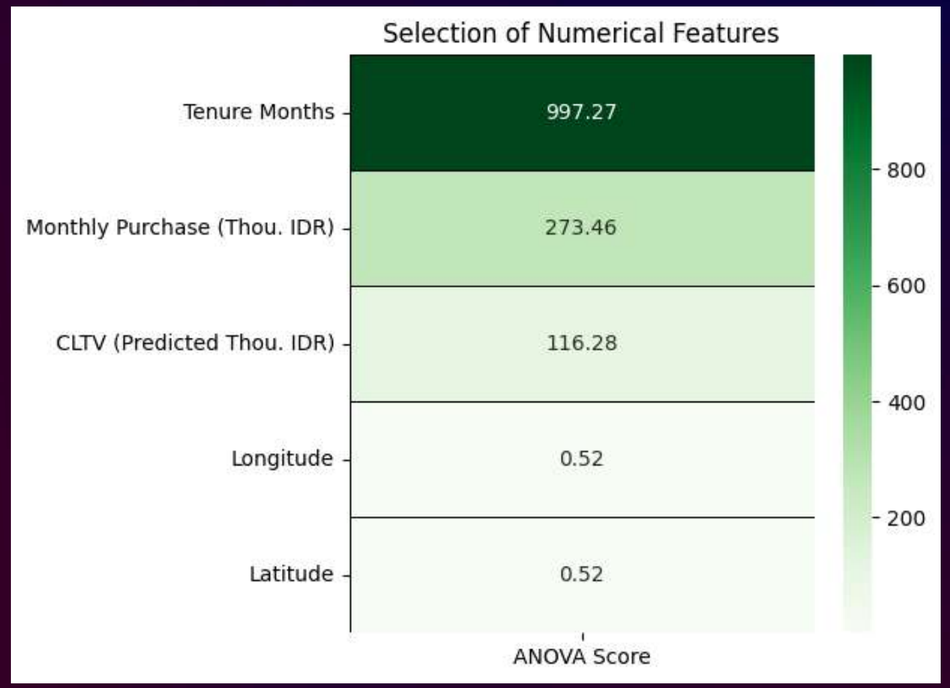
To determine the relationship between numerical features and target feature

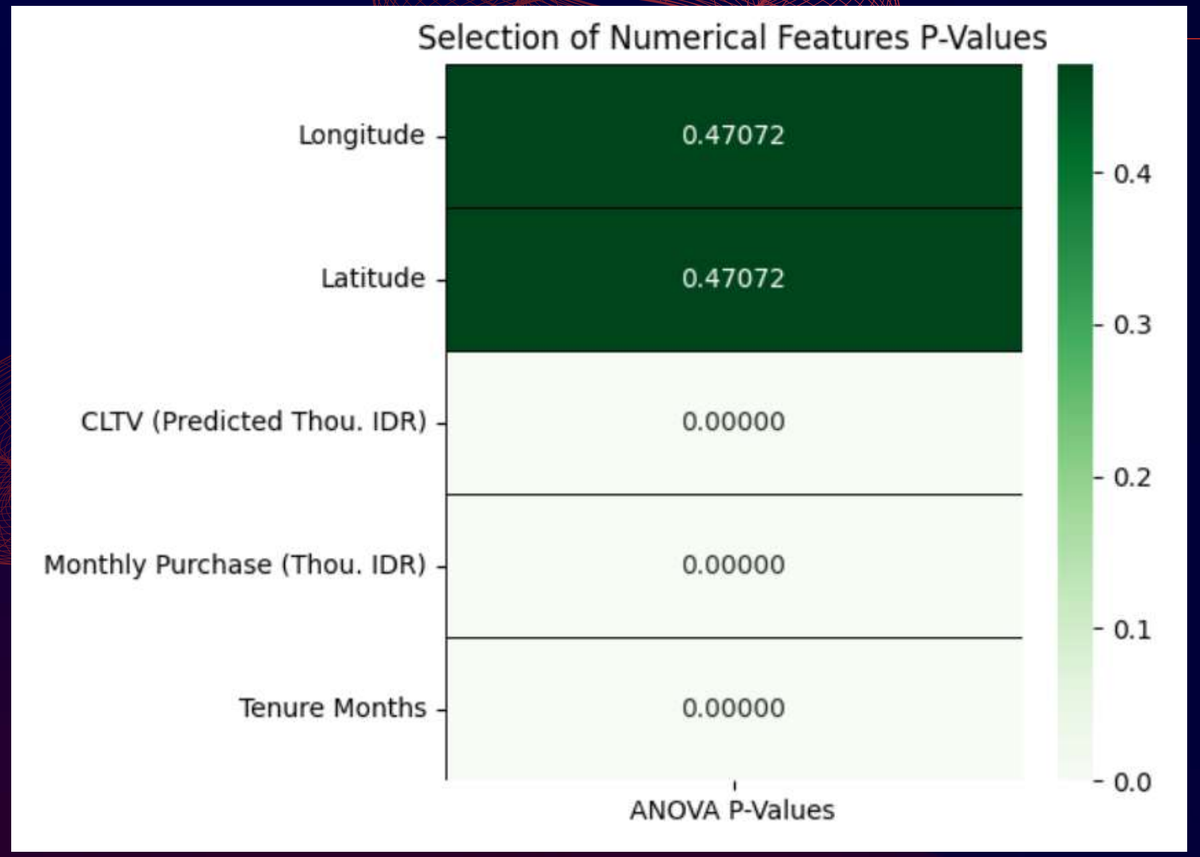
Define hypothesis

- Null Hypothesis (H0): There is no significance relationship between the numerical features and the target feature Churn Label.
- Alternate Hypothesis (H1): is that there is a significant relationship between the numerical features and the target feature Churn Label.

Feature Selection

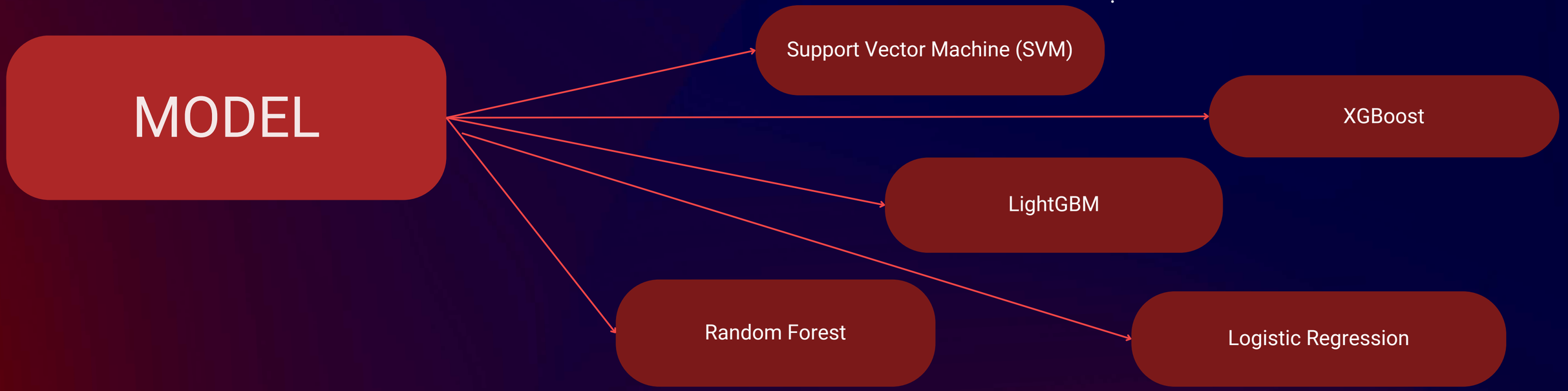
Chi-square





P-Values of Longitude and Latitude are greater than 0.05. So we can reject the null hypothesis. Then, we don't use them to feature for modeling

Model Building



After defining the models, the code evaluates their performance using cross-validation and various metrics, including accuracy, precision, recall, and F1-score.

Model Evaluation

Recall (Sensitivitas or True Positive Rate) : measures how well a model can identify all the actual positive instances.

$$\text{Recall} = 2\,x\,\frac{True\ Positive}{True\ Positive + False\ negative}$$

Precision (Positive Predictive Value) : indicates how well a model can identify positive instances and how often its predictions are correct.

$$\text{Precission} = 2\,x\,\frac{True\ Positive}{True\ Positive + False\ positive}$$

Evaluation Metric: F1-Score

- F1-Score is an evaluation metric that combines precision and recall. It is particularly useful when dealing with imbalanced classes.

$$F1 = 2\,x\,\frac{precision\ x\ recall}{precision + recall}$$

Handling Data Imbalance: SMOTE

- SMOTE is used to address class imbalance by generating synthetic samples of the minority class. SMOTE creates new samples in the feature space by combining existing data points.

SMOTE and Hyperparameter Tuning

Handling Data Imbalance: SMOTE

- SMOTE is used to address class imbalance by generating synthetic samples of the minority class. SMOTE creates new samples in the feature space by combining existing data points.

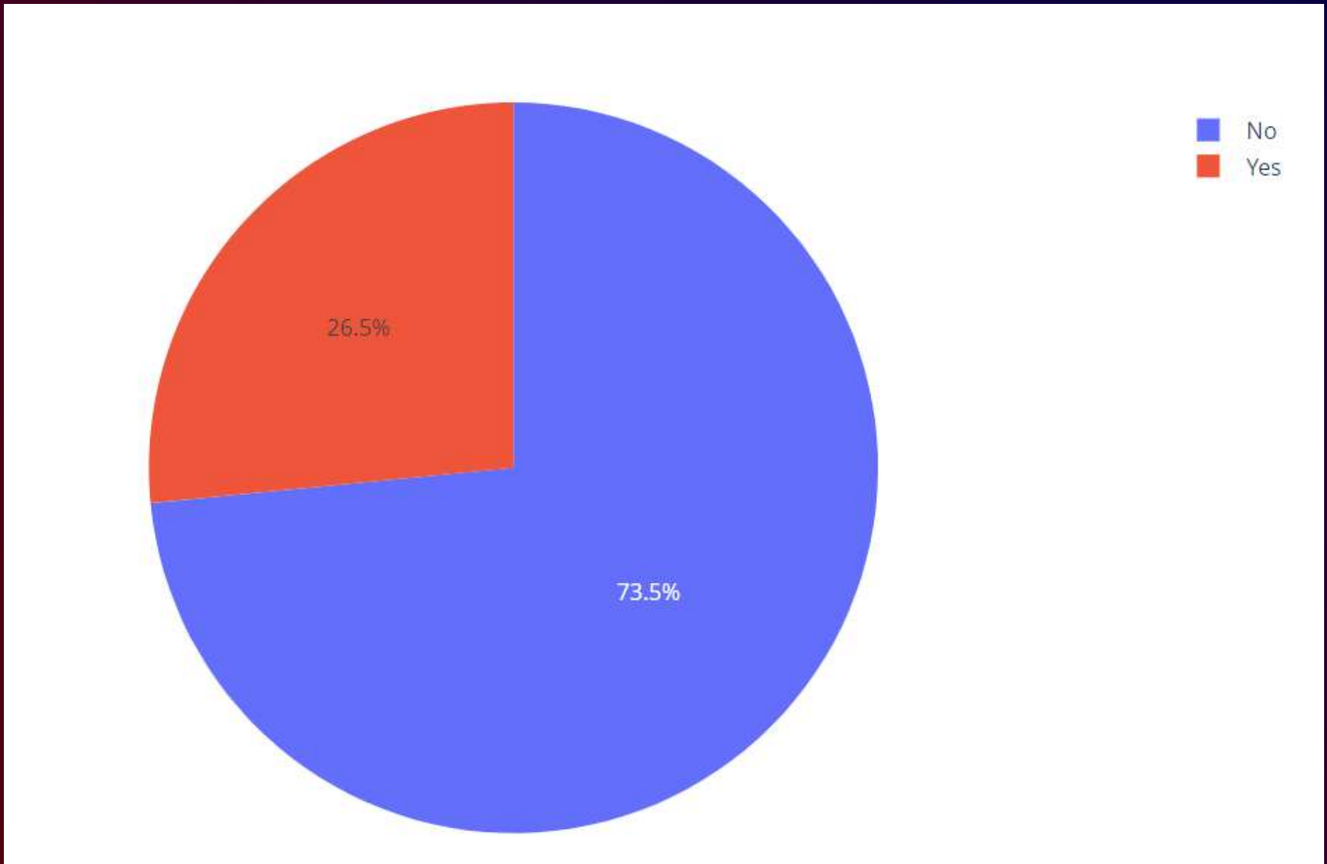
Hyperparameter Tuning

- Search best parameter for each model using RandomizedSearchCV

```
svc_params :  {'kernel': 'rbf', 'gamma': 1, 'C': 1000}
rf_params :  {'n_estimators': 130, 'max_depth': 17, 'criterion': 'entropy'}
xgb_params :  {'learning_rate': 0.057580335594252116, 'max_depth': 8, 'n_estimators': 91, 'subsample': 0.78735034326044
7}
lgbm_params :  {'num_leaves': 100, 'min_child_weight': 0.01, 'max_depth': 10, 'learning_rate': 0.1}
logreg_params :  {'penalty': 'l2', 'intercept_scaling': 2, 'C': 0.8}
```

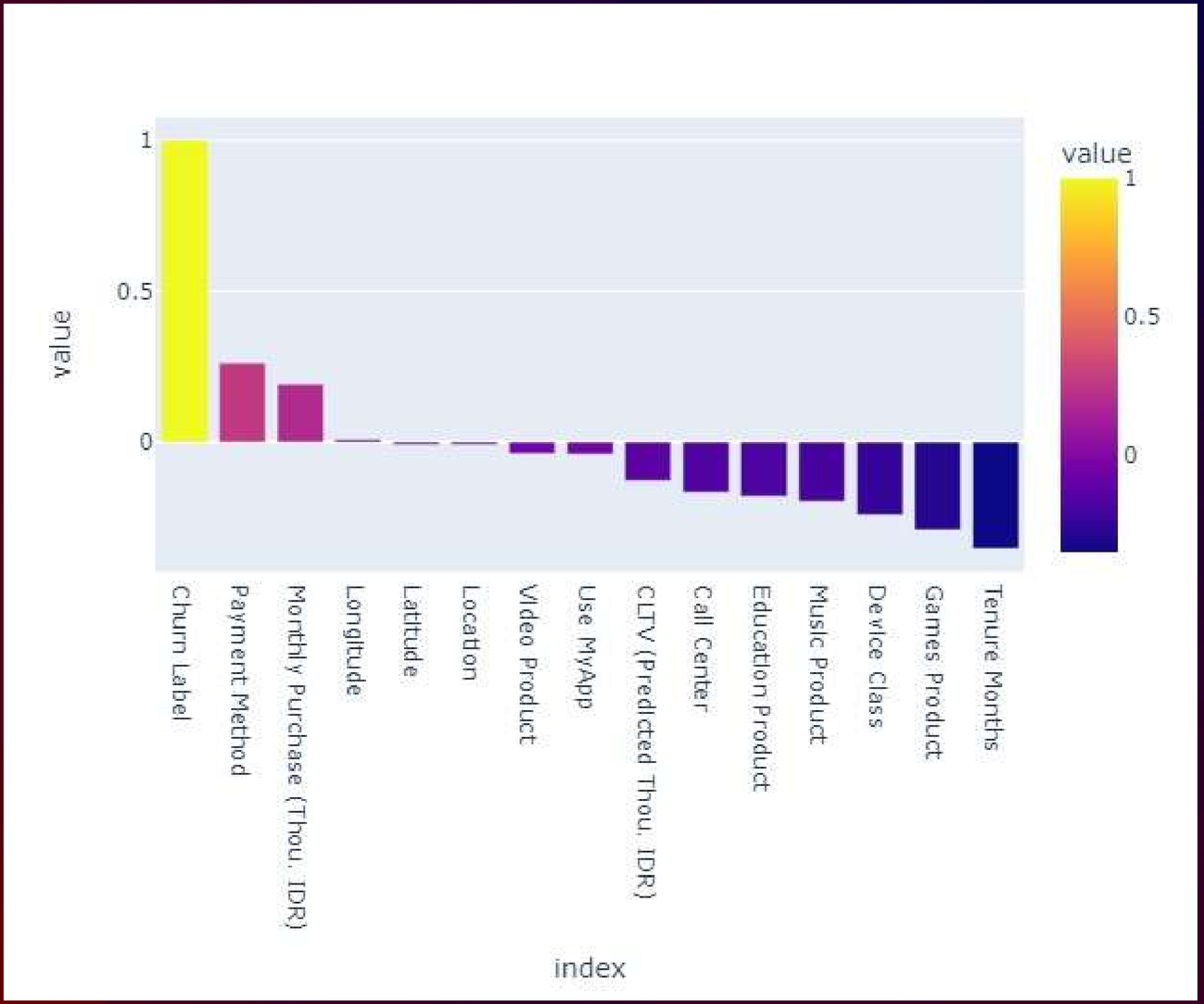
Results and Business Insight

Distribution of Target Feature



Customers who churn are 26.5% and those who don't churn are 73.5%. This is an unbalanced target feature.

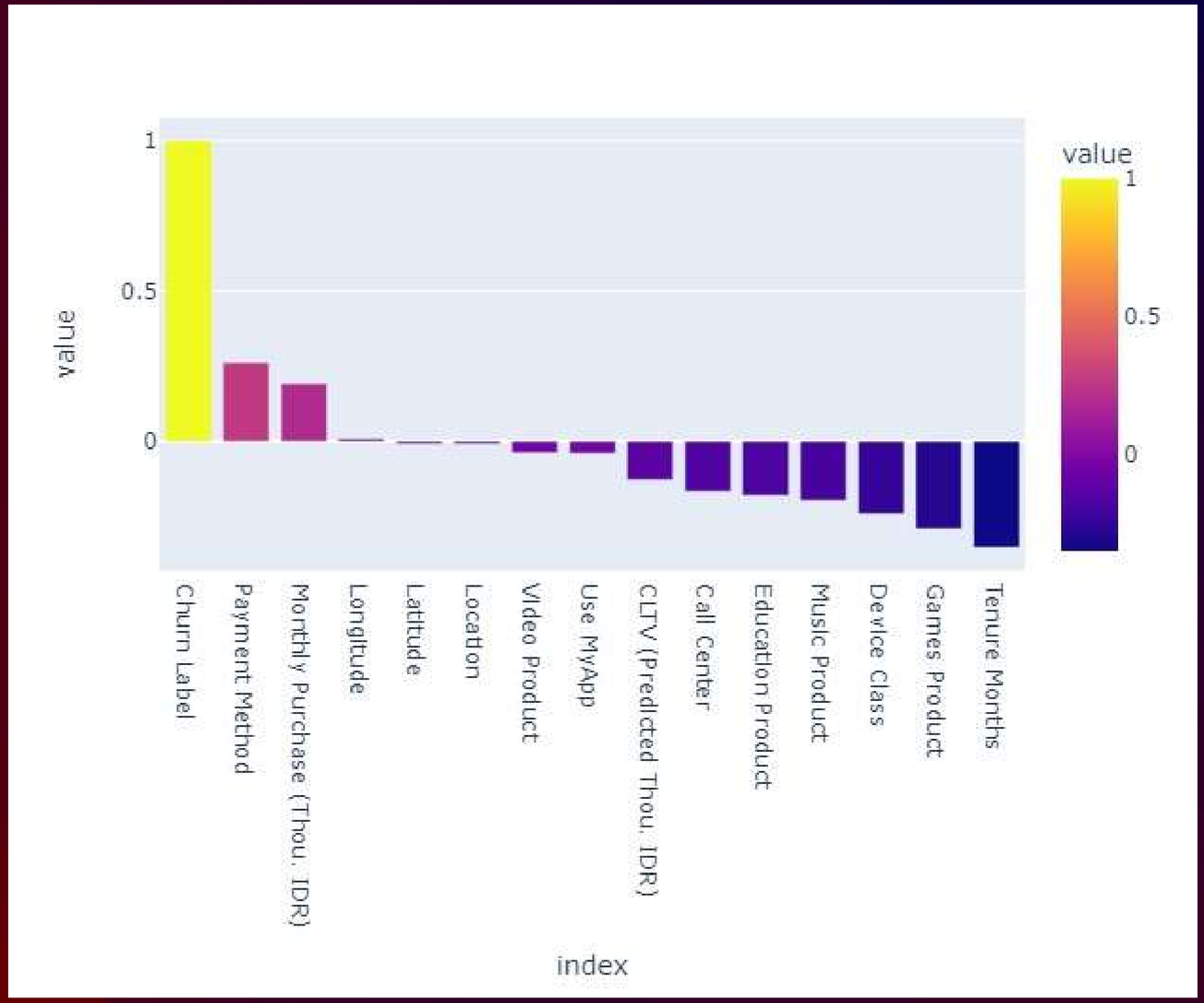
Correlation Analysis



In the results of data analysis using the correlation, it can be seen that all variables affect the tendency of customers to churn.

The bluer the color of the barplot, the greater the correlation with customers who do not churn, and the more yellow the color, the greater the correlation with customers who churn.

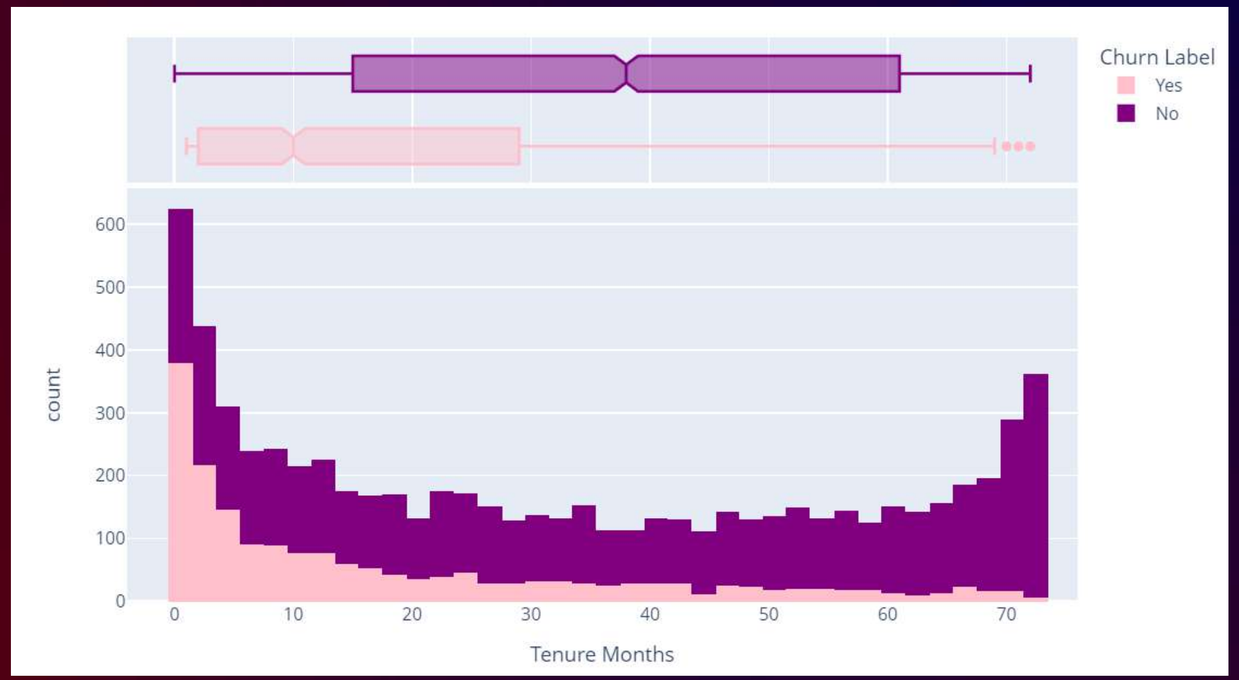
Correlation Analysis



Tenure Months, Games Product, and Device Class are the three most influential factors in customer retention.

Payment Method and Monthly Purchase are the most influential factors in customer churn.

Tenure Months by Churn

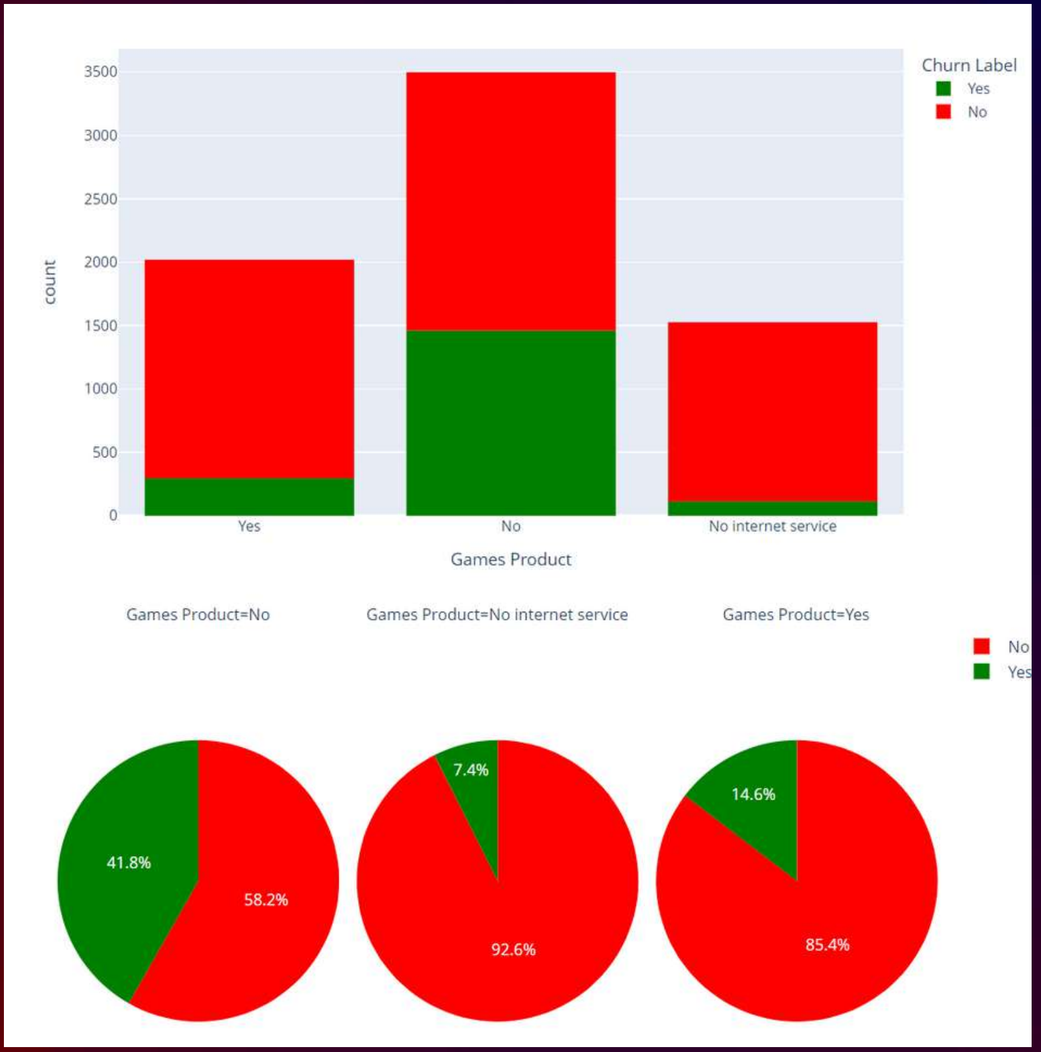



```
Churn Label
No           0.25    15.0
             0.50    38.0
             0.75    61.0
Yes          0.25     2.0
             0.50    10.0
             0.75    29.0
Name: Tenure Months, dtype: float64
```

Churn tends to occur in customers who have used the service recently. Even if we look at the median, **50% of customers who churn only use the service for less than 10 months.**

Customer with a low tenure months usually just take some new user discount or flash sales. **The company need a new strategy by giving them XP (poins) for every purchase or the app's coins for every cashback a customer obtain**. The company could also record customer's behaviour and give them vouchers based on their personal needs. These could rise customers retention and make the customers more valuable to the company.

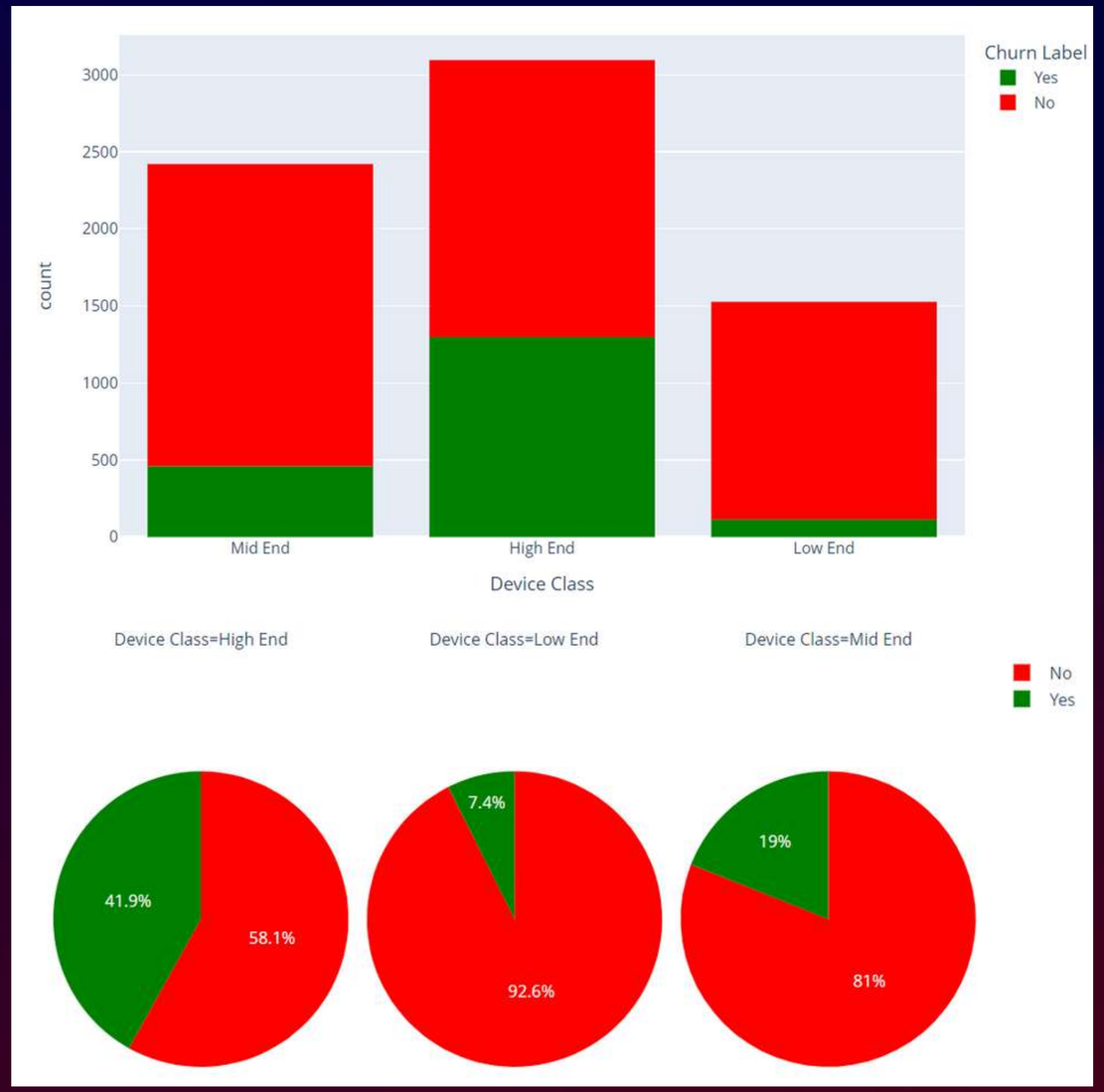
Games Product by Churn



Games product has a high influence on preventing customer to churn with or without the internet service. The company can use customer's behavior to look at what are the games that customers mostly use by looking at the history purchase. **The company can offer partnership with the games developer to make a great deal.** Therefore, the company can ensure customer's loyalty by making better offers than competitors.

Device Class by Churn

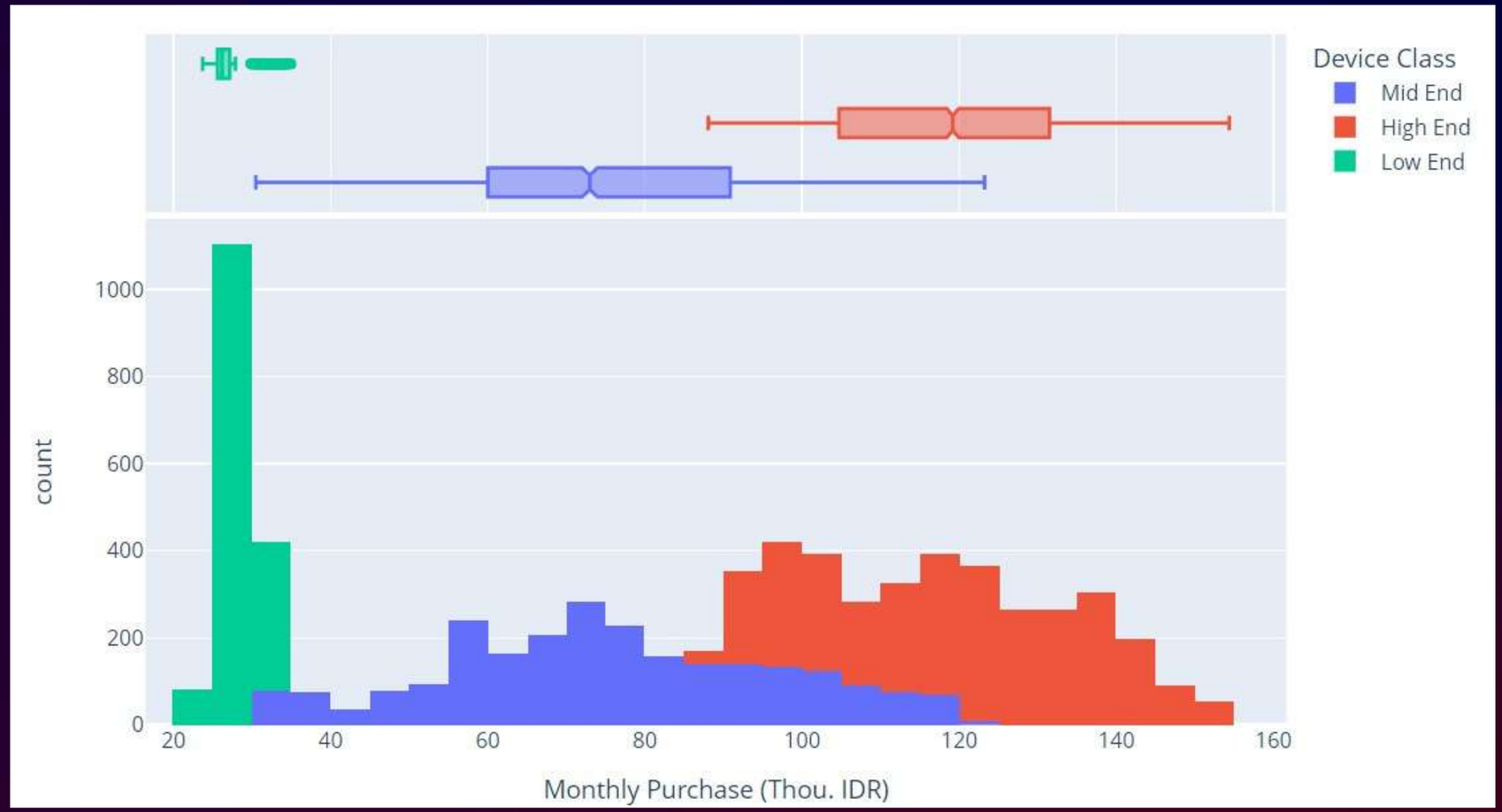
Device Class is the third categorical feature for its influence on customer churn. Based on this data, the mid-end and low-end device users have a tendency to be loyal 20-30% than high-end device users. **The company can focus maintain their strategy on low-end and mid-end device.**



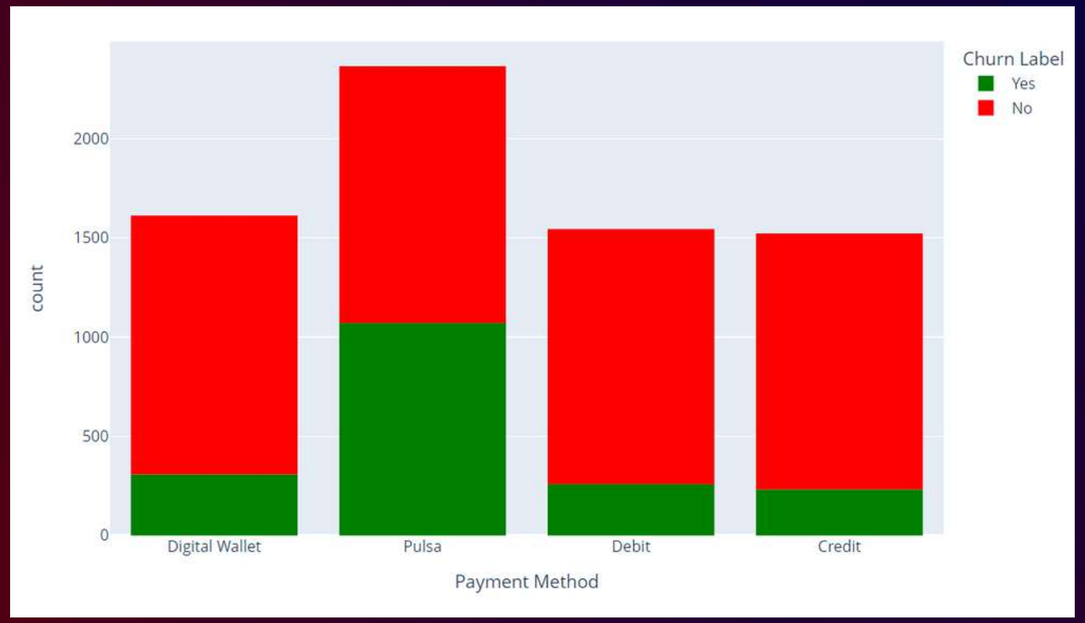
Monthly Purchase by Device Class

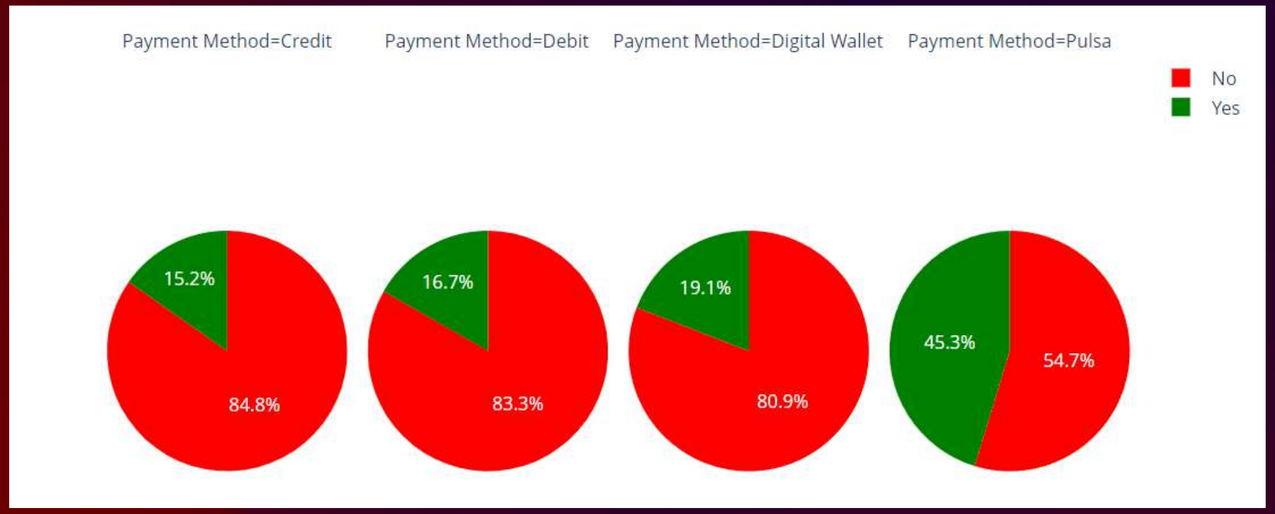
From the correlation beetween Monthly Purchase on Device Class we can see that the High-end device users spend more that the Mid-end and Low-end device users.

To rise customer retention on high-end device users, **the company can give more voucher or discount for high-end device.**



Payment Method by Churn





Payment Method has a significant influence to customer churn. **Based on the data, Digital wallet, Debit, and Credit users have more than 80% tendency not to churn**. All those variables have higher convinience and flexibility than using Pulsa as a payment method.

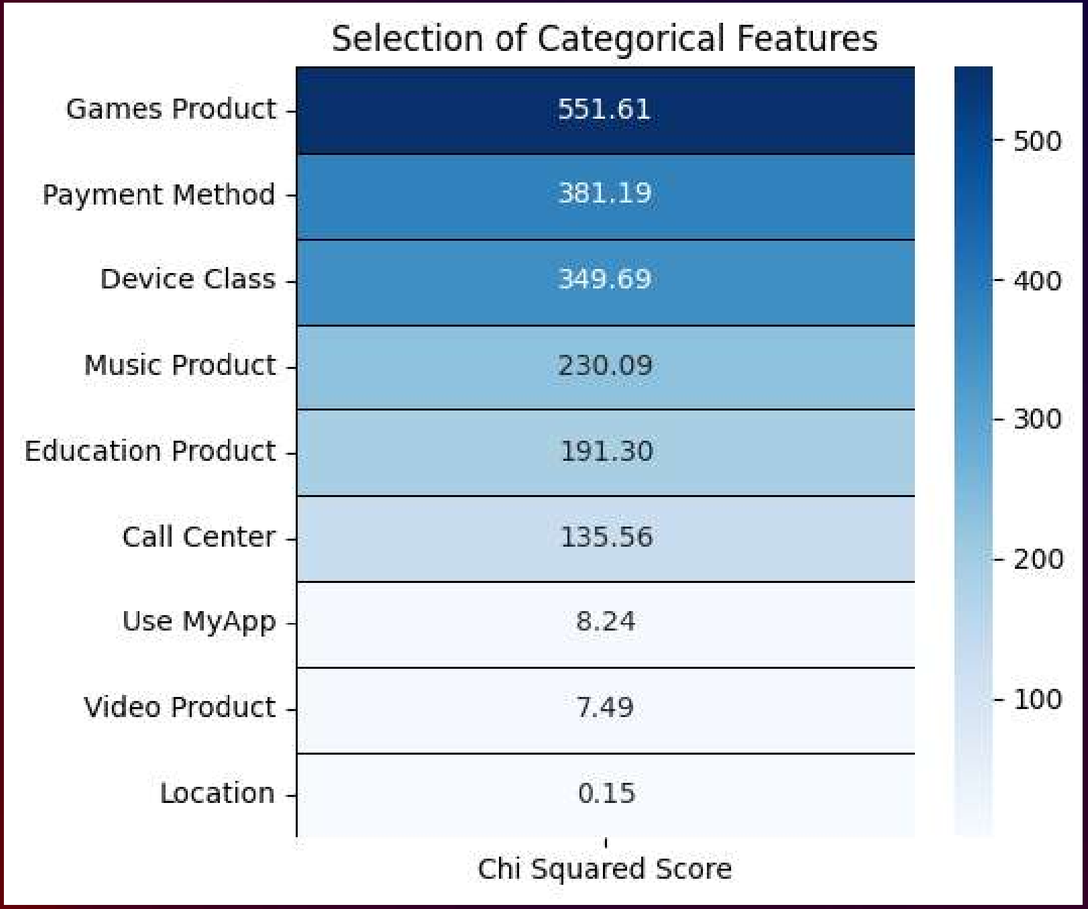
The company can improve their UI/UX design in order to rise customer retention. **The company can also offer a network loan just before a customer disconnected from the network. More than that, company can also start a monthly credit system.** By making it easier, customers will feel more served, helped, and most importantly emotionally fullfilled.

Monthly Purchase by Churn


```
Churn Label
No              0.25     32.6300
                0.50     83.7525
                0.75    114.9200
Yes             0.25     72.9950
                0.50    103.5450
                0.75    122.4600
Name: Monthly Purchase (Thou. IDR), dtype: float64
```

Based on the graph, customers with low monthly purchases have a tendency not to churn. To target the right customers, companies should focus on customer's monthly purchase that less than 40-50 (Thou. IDR) for the first quantile. These types of customers tend to be loyal and are the right targets for the company to provide more discounts or better promos and thus increase the company's revenue.

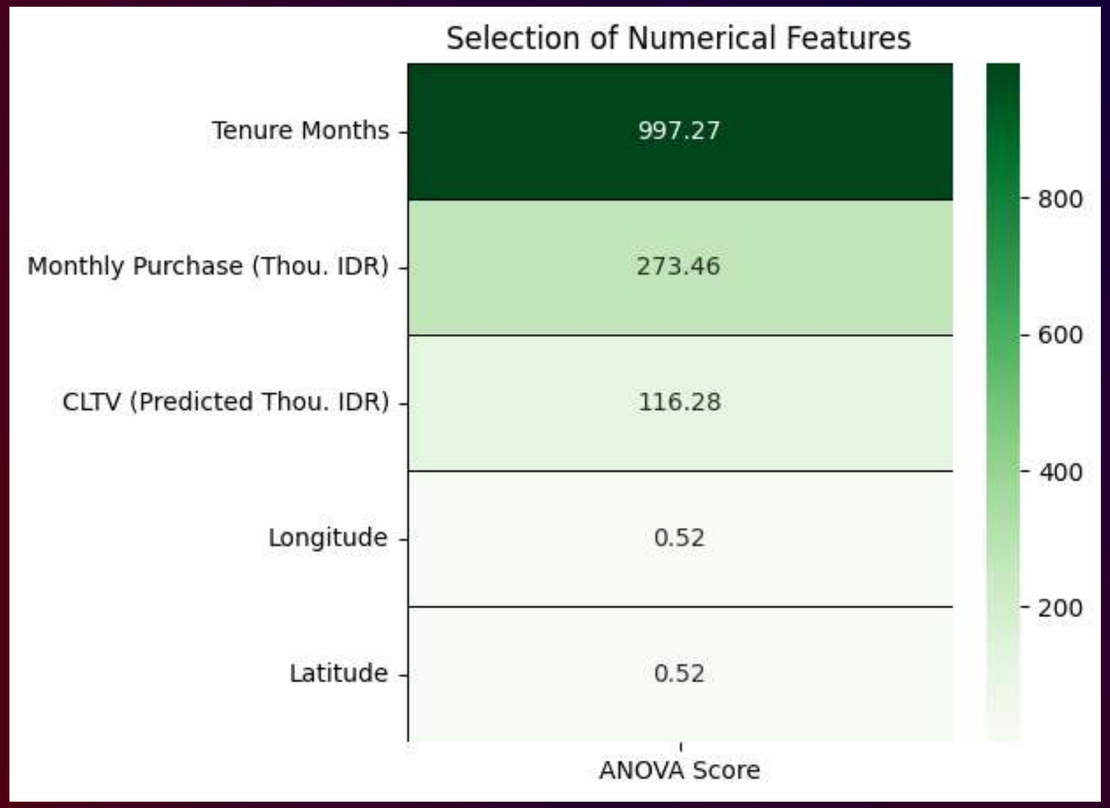
Data Analysis



The heatmap form of the Chi Squared Score displays the results of Feature Importance and Feature Selection in Categorical Features.

The results show that Games Product Payment Method and Device Class are three important factors in categorical features.

Data Analysis



The heatmap form of the ANOVA Score displays the results of Feature Importance and Feature Selection on Numerical Features.

The results obtained show that Tenure Months is very influential on the tendency of customers to churn, followed by Monthly Purchase and CLTV.

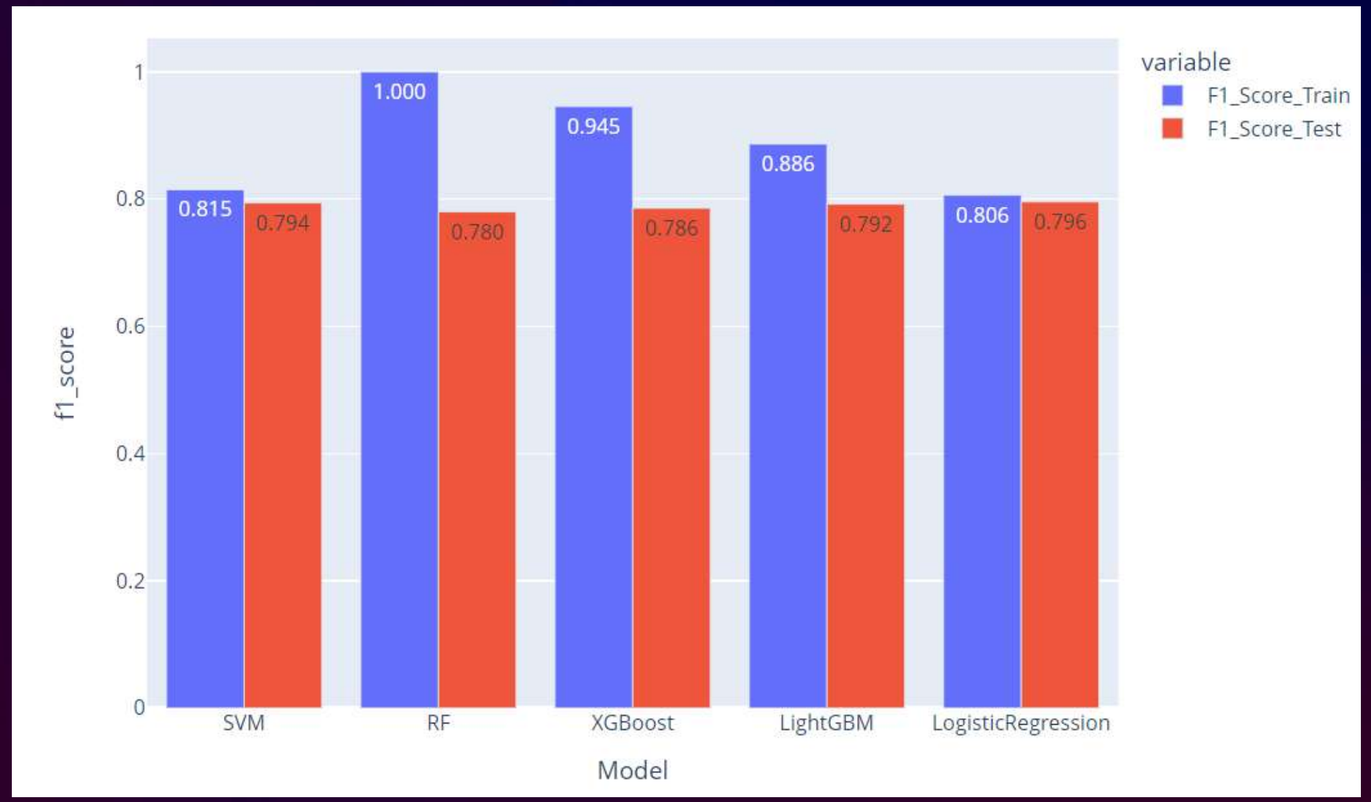
Results Modeling and Evaluation Model

Main Matrix F1-Scores

The evaluation model used is F1-Score because it wants to prevent False Negatives and False Positives

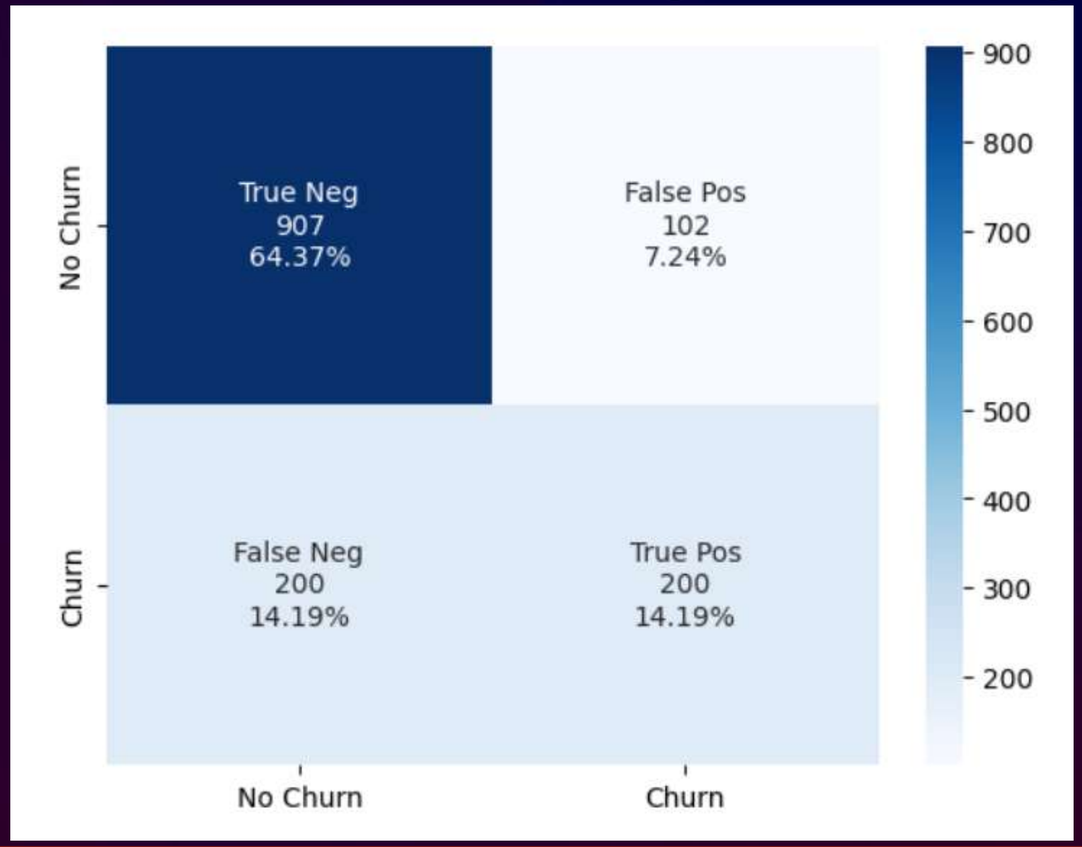
- If the FN (False Negative) has a large value, it will result in a decrease in the company's revenue.
- If the FN (False Negative) is large, then efforts to find customers/customer acquisition will require greater operational costs than retaining old customers.
- If FP (False Positive) has a large value, it means that there is an error in predicting customers who are not actually churning, but our model predicts churn.
- If the FP (False Positive) is large, then it is said that the analyst mistargeted in placing operational costs on customer retention, which should be used for other operations such as campaigns. The lower FP value is, the higher chance for a company to save their costs.

F1-Score Analysis



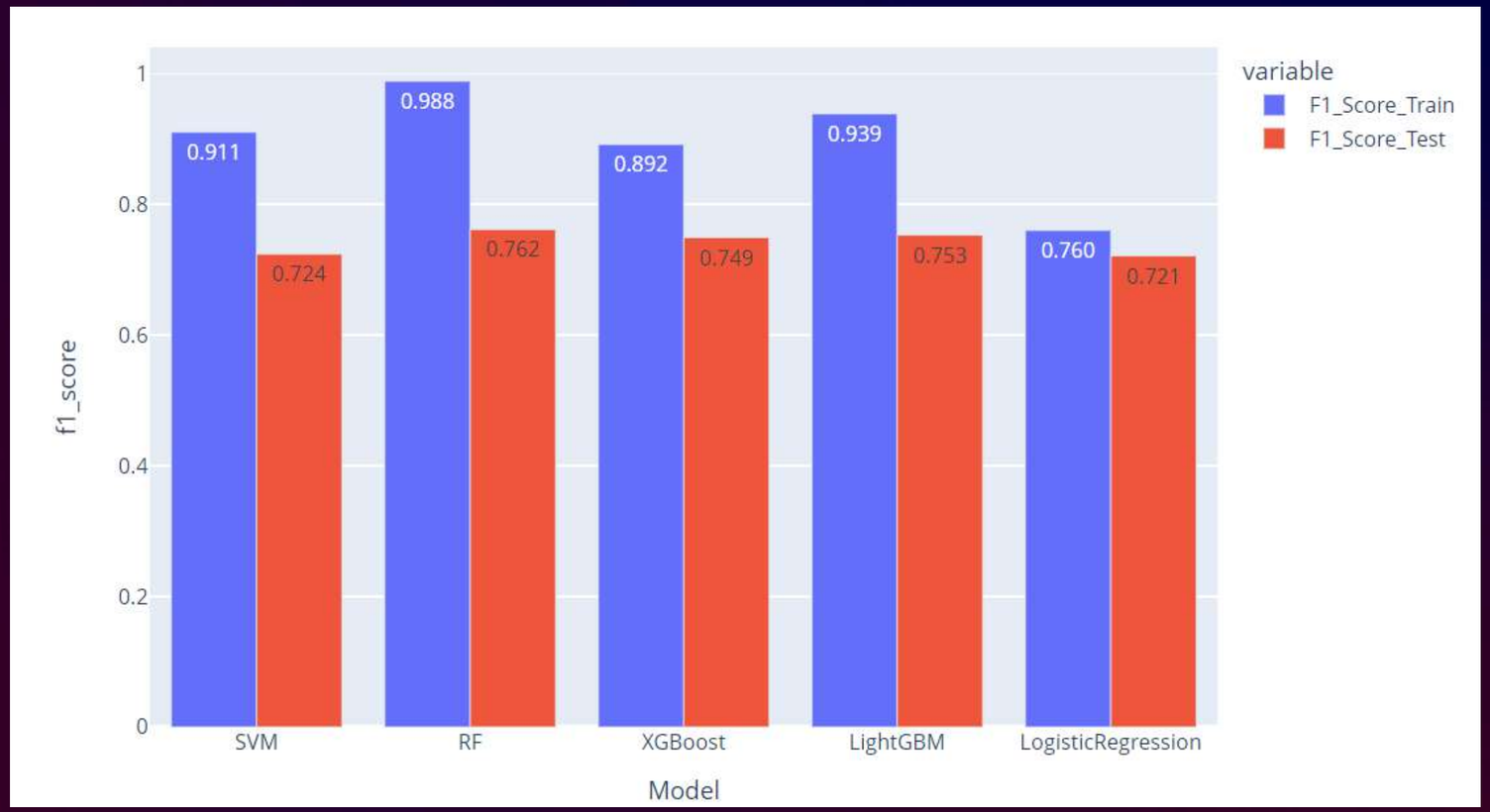
Random Forest and XGBoost are overfitting because the train and test f1 scores are far different.
Then, a good model that is not overfitting and has the highest f1 score is Logistic Regression.

Confusion Matrix from Logistic Regression



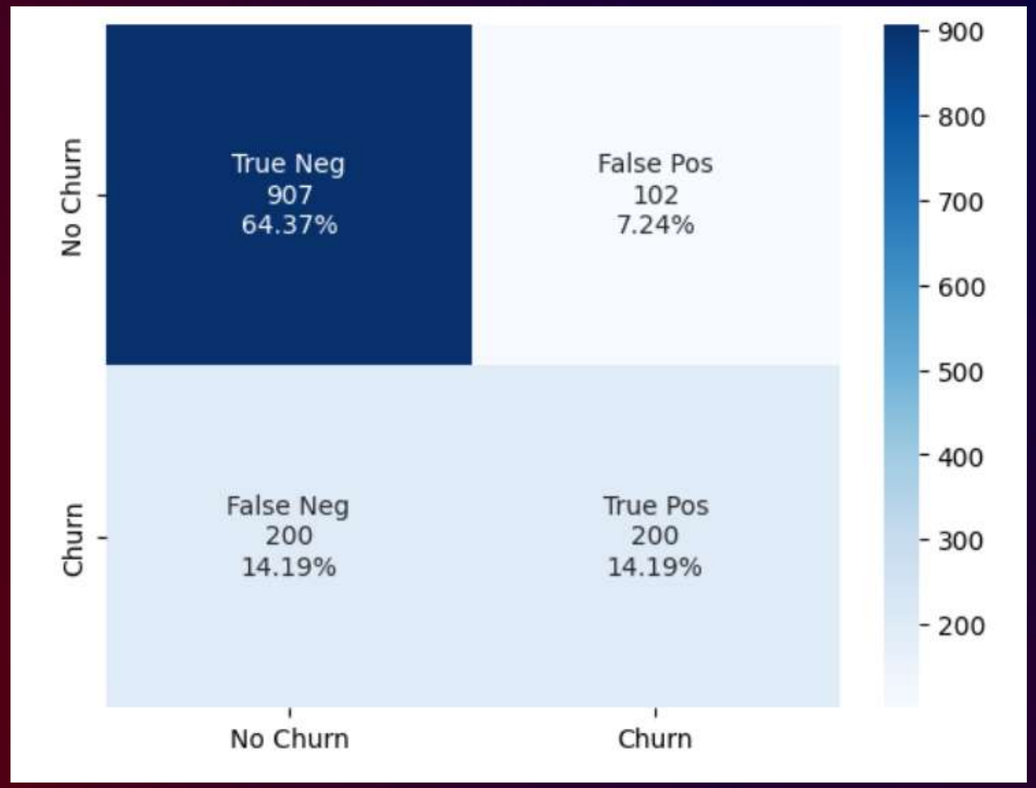
This Logistic Regression model has False Negative greater than False Positive so this model better for prevent operational cost to customer retention.

F1-Score Analysis for SMOTE and Hyperparameter Tuning



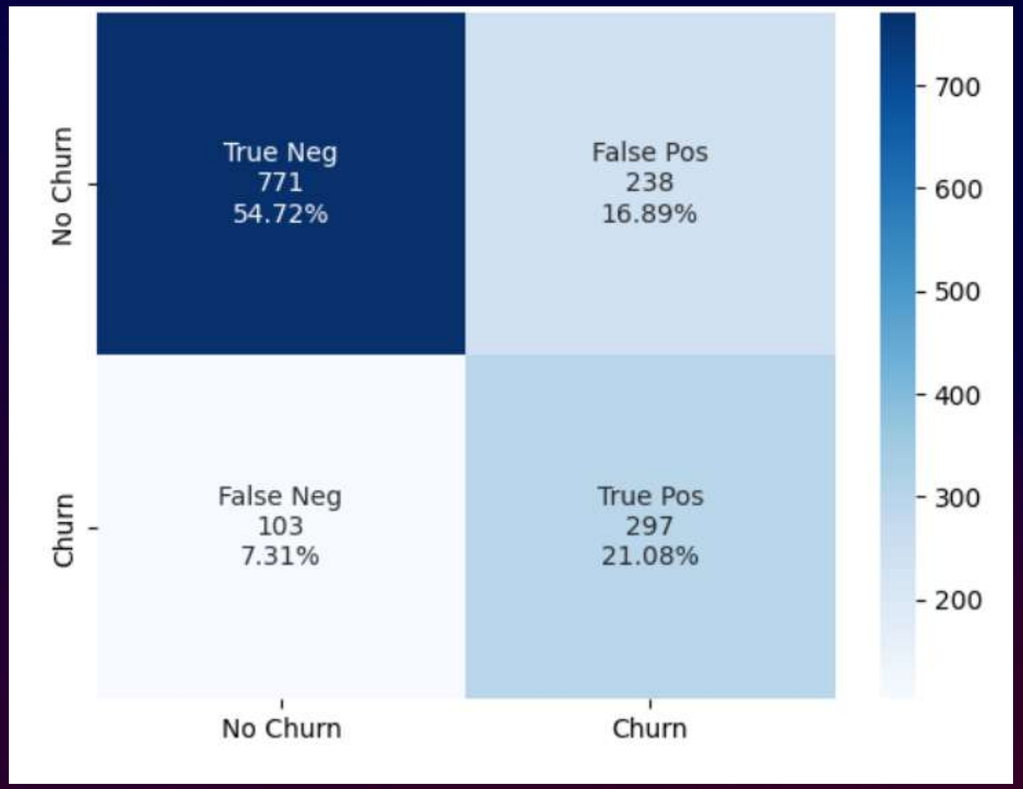
The best f1-score for data testing is Random Forest model, but that look overfitting. XGBoost is the second best model. Model does not overfitting is Logistic Regression, but has a f1-score less then all models.

Confusion Matrix



Random Forest Classifier

This model prevents operational costs due to False Positive being greater than False Negative.



XGBoost Classifier

This model prevents the company's revenue from being reduced because the False Negative is greater than the False Positive.

Conclusion

Telco Churn-Data Science Indonesia

The majority of customers who use cashback are customers with less than 10 months of service usage. Here are some factors or customer activities that cause churn, among others:

1. Tenure Months has the highest correlation with customer churn
2. Games Product is quite significant in increasing customer retention.
3. Flexible and fast payment methods are much preferred by customers.
4. Monthly Purchase customers who are few can characterize customers who must be retained.

The best model used to prevent overvitting and has the highest F1 Score value is **Logistic Regression**. The confusion matrix value in Logistic Regression has a False Positive value that is smaller than the False Negative so that this model is suitable for avoiding excessive use of company operational costs that can be used inappropriately. **A larger False Negative value also indicates that attracting new customers requires more operational costs than maintaining existing customers.**

Thank You