Classification Model with 93% Accuracy – Telco Customer Churn Analysis

12 Jan 2024

by Ahmad zaki bin ramli

Table of Content

Self Introduction	- Education, Work experience & About me
Course Timeline	- The period spanning from month 1 to month 6
Technology Stack	- Language, tools and environment
Project Overview, Problem Statement & Objective	- This project aims to address industry problems by achieving objectives through the defined step
Exploratory Data Analysis	- Data Inspection, Anamolies Detection (Distribution & Boxplot)
Feature Impotance (Random Forest)	- Encoding, Random Forest ML, Hyperparameter Tuning & Top of feature
Predictive Modeling	- Encoding, Machine learning, Hyperparameter Tuning & Result
Solution	- Recommendations for reducing churn
Project Timeline	- Milestones and key deliverables
Project Budget	- Estimated cost breakdown

Self Introduction

Universiti Kebangsaan Malaysia
The National University of Malaysia

- **About me**
  - **Name**: Zaki
  - **Hometown**: Ipoh, Perak.
  - **Current Location**: Living in Putrajaya for work purposes
  - **Passion:** Determined to explore new fields and continuously learn and improve.

- **Education**
  - **Degree**: Bachelor's in Economics
  - **Graduate Year**: End of 2022
  - **Challenges**: Completed degree during the COVID-19 pandemic, which made learning and exams difficult due to the shift to online platforms
  - **Achievement:** Despite challenges, remained focused on pursuing personal growth and career development.

- **Work experience**
  - **Position**: Personal MYSTEP (Graduate Program) at the Ministry of Economic Affairs
  - **Duration**: Almost 2 Years
  - **Responsibilities**: Gained experience in management and government-related work, assisting in economic projects and initiatives.

MYSTEP

Course Timeline

Month 1

- Self resilience mastery, Communication Mastery, career Launcpad Mastery

Month 2

- Introduction of Data Science
- Set up Environment
- Introduction to Language python and Others Libraries (Pandas)

Month 3

- Introduction to Numpy
- Data Visualization with Seaborn
- Web Scraping
- Introduction Language SQL

Month 4

- Introduction Statistical Data Analysis
- Introduction to Machine Learning
- Supervised Learning
- Unsupervised Learning
- Introduction to Recommended System

Month 5

- Introduction to Deep Learning & NLP
- Capstone Project
- Self Leadership

Month 6

- Final Review

Technology Stack

- **Data Analysis Tools**
- **Visualization Libraries**
- **Machine Learning Frameworks**

Project Overview

- **Customer churn impacts the revenue and growth of telecommunications companies** like **Celcom**.
- **Celcom needs to identify the key factors** that affect customer churn, such as pricing, network quality, customer service, and service offerings.
- **This project aims to predict customer churn** by analyzing data on Celcom's customers and their service usage patterns.
- **Identifying patterns and trends in the data** will help Celcom improve its customer retention strategies, allowing them to take actions to prevent churn and retain valuable customers.

Problem Statement

- **Customer churn directly affects the profit and growth of telecommunications companies** like Celcom.
- **In a competitive market, companies** such as Celcom, Maxis, Digi, and U Mobile face chall
- **Customers may leave due to factors** such as pricing, service quality, customer support, or better offers from competitors.
- **It is important to identify the key factors influencing churn**, such as network reliability, customer service satisfaction, and pricing structure.
- **Predicting the likelihood of churn helps companies** like Celcom take effective preventive actions to retain customers.

Objective

- Identify the **factors influencing** churn and patterns in the data.
- **Predict customer churn** using machine learning model.
- **Help Celcom companies improve** their customer retention strategies.

Exploratory Data Analysis

(20) **X = FEATURE**

- CustomerID
- Gender
- SeniorCitizen
- Partner
- Tenure
- Dependents
- MultipleLines
- InternetService
- OnlineSecurity
- OnlineBackup
- DeviceProtection
- Techsupport
- StreamingTV
- StreamingMovies
- Contract
- PaperlessBilling
- PaymentMethod
- PhoneService
- MonthlyCharges
- TotalCharges

(1) **Y = TARGET**

- Churn

```
df['Churn'].value_counts()
```

```
Churn
No     5174
Yes    1869
Name: count, dtype: int64
```

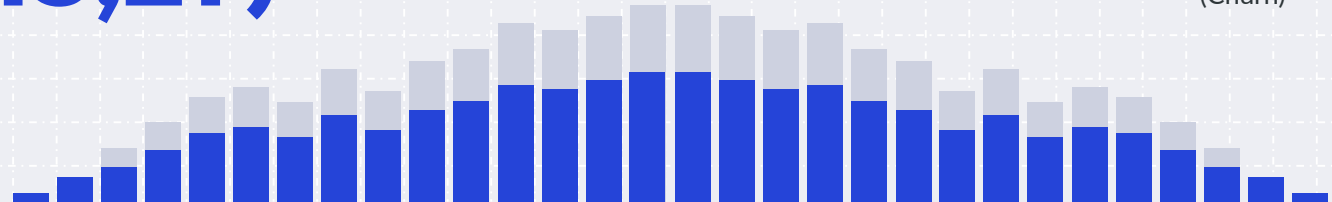
```
# Get the percentage of each unique value in the 'Attrition' column
Churn_percentage = df['Churn'].value_counts(normalize=True) * 100

# Print the result
print(Churn_percentage )
```

```
Churn
No     73.463013
Yes    26.536987
Name: proportion, dtype: float64
```

- Shown imbalance dataset (Churn)

(7043,21)



Exploratory Data Analysis

Inspection Datatypes, Missing Values

checking the following parts of the data for an initial review :

- ➔ Datatypes
- ➔ Duplicates data
- ➔ Missing values

```
df.shape
```

```
(7043, 21)
```

```
df.nunique()
```

```
customerID          7043
gender                 2
SeniorCitizen          2
Partner                2
Dependents             2
tenure                73
PhoneService           2
MultipleLines          3
InternetService        3
OnlineSecurity         3
OnlineBackup           3
DeviceProtection       3
TechSupport            3
StreamingTV            3
StreamingMovies        3
Contract               3
PaperlessBilling       2
PaymentMethod          4
MonthlyCharges      1585
TotalCharges        6531
Churn                  2
dtype: int64
```

- Don't have any duplicated data

```
customerID          object
gender              object
SeniorCitizen        int64
Partner             object
Dependents          object
tenure               int64
PhoneService        object
MultipleLines       object
InternetService     object
OnlineSecurity      object
OnlineBackup        object
DeviceProtection    object
TechSupport         object
StreamingTV         object
StreamingMovies     object
Contract            object
PaperlessBilling    object
PaymentMethod       object
MonthlyCharges     float64
TotalCharges        object
Churn               object
dtype: object
```

- Need to change to numerical instead of object

```
df.isna().sum()
```

```
customerID          0
gender              0
SeniorCitizen       0
Partner             0
Dependents          0
tenure              0
PhoneService        0
MultipleLines       0
InternetService     0
OnlineSecurity      0
OnlineBackup        0
DeviceProtection    0
TechSupport         0
StreamingTV         0
StreamingMovies     0
Contract            0
PaperlessBilling    0
PaymentMethod       0
MonthlyCharges      0
TotalCharges        0
Churn               0
dtype: int64
```

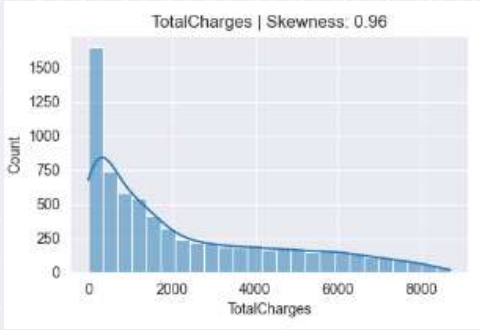
- Don't have any missing values

Exploratory Data Analysis

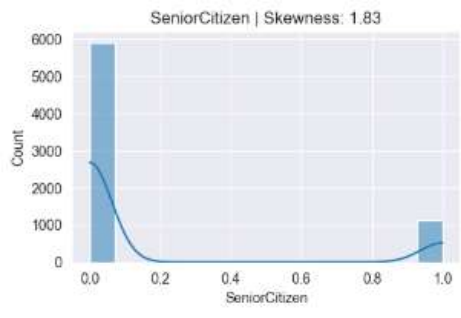
	SeniorCitizen	tenure	MonthlyCharges	TotalCharges
count	7043.000000	7043.000000	7043.000000	7043.000000
mean	0.162147	32.371149	64.761692	2279.734304
std	0.368612	24.559481	30.090047	2266.794470
min	0.000000	0.000000	18.250000	0.000000
25%	0.000000	9.000000	35.500000	398.550000
50%	0.000000	29.000000	70.350000	1394.550000
75%	0.000000	55.000000	89.850000	3786.600000
max	1.000000	72.000000	118.750000	8684.800000

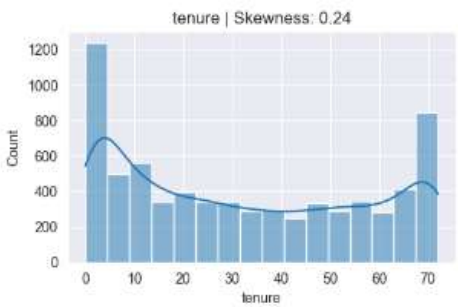
- Basic statistic

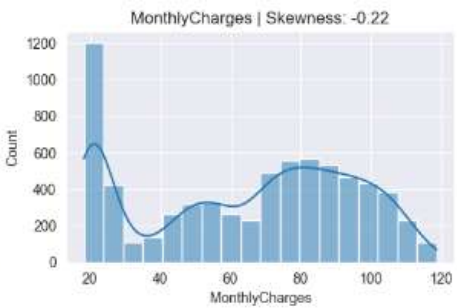
Anamolies Detection (Stats / Distribution)



- Total Charge looks suspicious at the first inspection

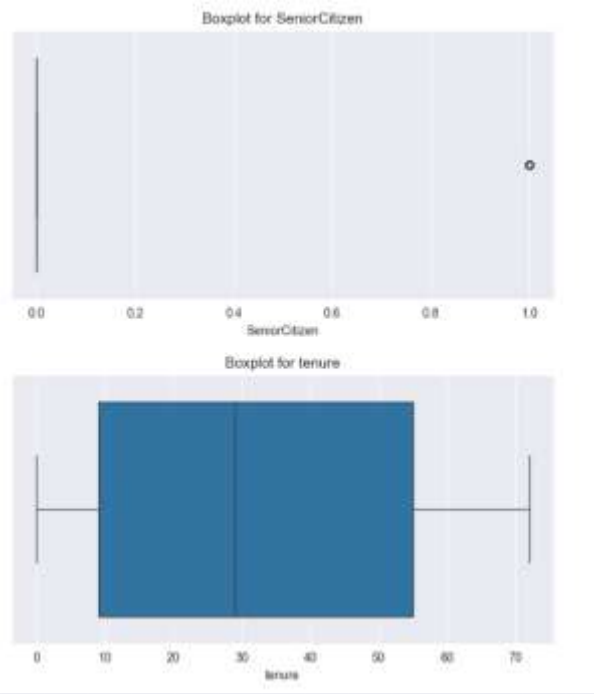


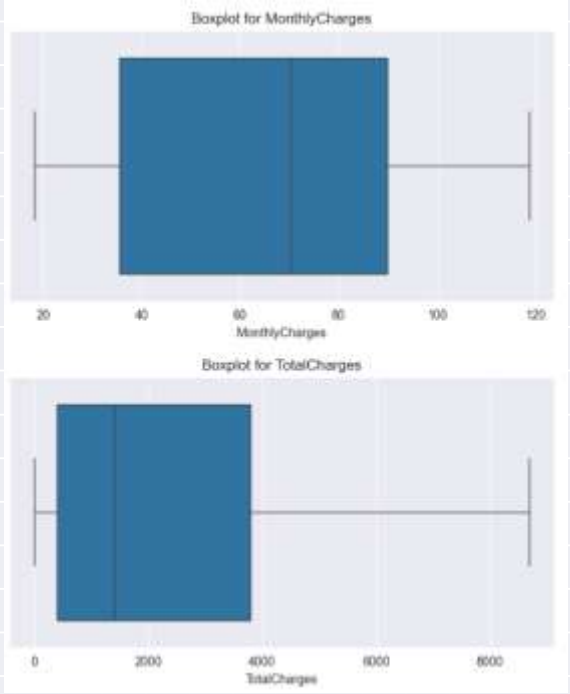




Exploratory Data Analysis

Anamolies Detection (Boxplot)





MonthlyCharges	TotalCharges
52.55	
20.25	
80.85	
25.75	
56.05	
19.85	
25.35	
20.00	
19.70	
73.35	
61.90	

Explanation

- Total Charge looks suspicious for the first inspection.
- Total Charge = 0 meaning they are a new customer and its really make sense and this is not the outliers
- Tenure & MonthlyCharges looks normal

Feature Importance

- **Data Transformation**
  - Using Label Encoding
  - Scaling (Standardscaler)

```
df_scaled.head()
```

	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup
0	0	0	1	0	-1.277445	0	1	0	0	2
1	1	0	0	0	0.066327	1	0	0	2	0
2	1	0	0	0	-1.236724	1	0	0	2	2
3	1	0	0	0	0.514251	0	1	0	2	0
4	0	0	0	0	-1.236724	1	0	1	0	0

- **Machine Learning (Random Forest)**

Model Overfitting

TRAIN

TEST

```
=== Performance on Train Set ===
Accuracy (Train): 0.9980475683351083

Classification Report (Train):
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      4139
           1       1.00      1.00      1.00      1495

    accuracy                           1.00      5634
   macro avg       1.00      1.00      1.00      5634
weighted avg       1.00      1.00      1.00      5634


--------------------------------------------------------

=== Performance on Test Set ===
Accuracy (Test): 0.7892112872391767

Classification Report (Test):
              precision    recall  f1-score   support

           0       0.83      0.90      0.86      1035
           1       0.63      0.48      0.55       374

    accuracy                           0.79      1409
   macro avg       0.73      0.70      0.71      1409
weighted avg       0.78      0.79      0.78      1409
```

After Tuning

Model Balance

```
Fitting 5 folds for each of 72 candidates, totalling 360 fits
Best Parameters: {'max_depth': None, 'min_samples_leaf': 5, 'min_samples_split': 15, 'n_estimators': 50}

Cross-Validation Scores: [0.79148181 0.7905945  0.78527863 0.78172138 0.74333925]
Mean CV Accuracy: 0.7784815153955628

=== Performance on Train Set ===
Accuracy (Train): 0.8485977990770323

Classification Report (Train):
              precision    recall  f1-score   support

           0       0.95      0.84      0.89      4139
           1       0.66      0.87      0.75      1495

    accuracy                           0.85      5634
   macro avg       0.81      0.85      0.82      5634
weighted avg       0.87      0.85      0.85      5634


=== Performance on Test Set ===
Accuracy (Test): 0.7643718949609652

Classification Report (Test):
              precision    recall  f1-score   support

           0       0.89      0.78      0.83      1035
           1       0.54      0.73      0.62       374

    accuracy                           0.76      1409
   macro avg       0.72      0.75      0.73      1409
weighted avg       0.80      0.76      0.77      1409
```

- This model ready for feature importance

Feature Importance

- **Hyperparameter Tuning**

  - Class Weight (Imbalance dataset)
  - Grid Search
  - Cross Validation

```
rf = RandomForestClassifier(random_state=42, class_weight="balanced")

# Hyperparameter grid
param_grid = {
    'n_estimators': [50, 100, 150],
    'max_depth': [None, 5, 10, 15],
    'min_samples_split': [5, 10, 15],
    'min_samples_leaf': [5, 10]
}


grid_search = GridSearchCV(estimator=rf, param_grid=param_grid,
                           cv=5, scoring='accuracy', verbose=2, n_jobs=-1)
```

```
cv_scores = cross_val_score(best_model, X_train, y_train, cv=5, scoring='accuracy')
print("\nCross-Validation Scores:", cv_scores)
print("Mean CV Accuracy:", cv_scores.mean())
```

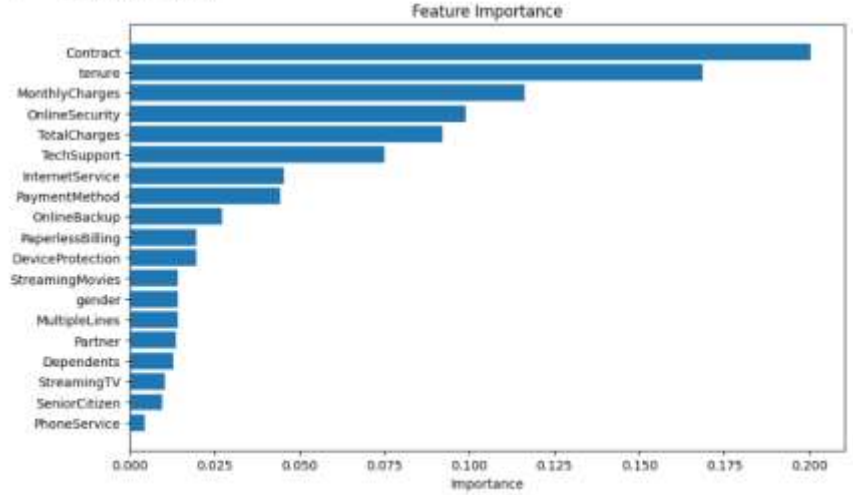
- **Top Feature Impotance**

Using Threshold Mean Feature Importance (Feature selection)

```
=== Feature Importance ===
            Feature  Importance
14         Contract    0.200376
4            tenure    0.168621
17   MonthlyCharges    0.116200
8    OnlineSecurity    0.098771
18    TotalCharges     0.092103
11     TechSupport     0.074966
7   InternetService    0.045346
10   PaymentMethod     0.044428
9     OnlineBackup     0.027075
15  PaperlessBilling  0.018543
10  DeviceProtection  0.018403
13  StreamingMovies   0.014229
0            gender   0.014044
6     MultipleLines   0.014024
2          Partner    0.013508
3       Dependents    0.012753
12     StreamingTV    0.010197
1     SeniorCitizen   0.009668
5     PhoneService    0.004475
```

- **Filtering process**

```
Selected Features (Importance > Mean):
            Feature  Importance
14         Contract    0.200376
4            tenure    0.168621
17   MonthlyCharges    0.116200
8    OnlineSecurity    0.098771
18    TotalCharges     0.092103
11     TechSupport     0.074966
```



Predictive Modeling

Data Transformation Using Top Feature

	Contract	tenure	MonthlyCharges	OnlineSecurity	TotalCharges	TechSupport	Churn
0	Month-to-month	1	29.85	No	29.85	No	No
1	One year	34	56.95	Yes	1889.5	No	No
2	Month-to-month	2	53.85	Yes	108.15	No	Yes
3	One year	45	42.30	Yes	1840.75	Yes	No
4	Month-to-month	2	70.70	No	151.65	No	Yes

One Hot Encoding

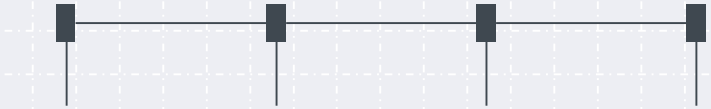
	tenure	MonthlyCharges	TotalCharges	Churn	Contract_Month-to-month	Contract_One year	Contract_Two year	OnlineSecurity_No	OnlineSecurity_No internet service	OnlineSecurity_Yes	TechSup
0	-1.277445	-1.160323	-0.992611	0	True	False	False	True	False	False	
1	0.066327	-0.259629	-0.172165	0	False	True	False	False	False	True	
2	-1.236724	-0.362660	-0.958066	1	True	False	False	False	False	True	
3	0.514251	-0.748535	-0.193672	0	False	True	False	False	False	True	
4	-1.236724	0.197365	-0.938874	1	True	False	False	True	False	False	

```
df_Filtered1.shape
```

```
(7043, 7)
```

Machine Learning

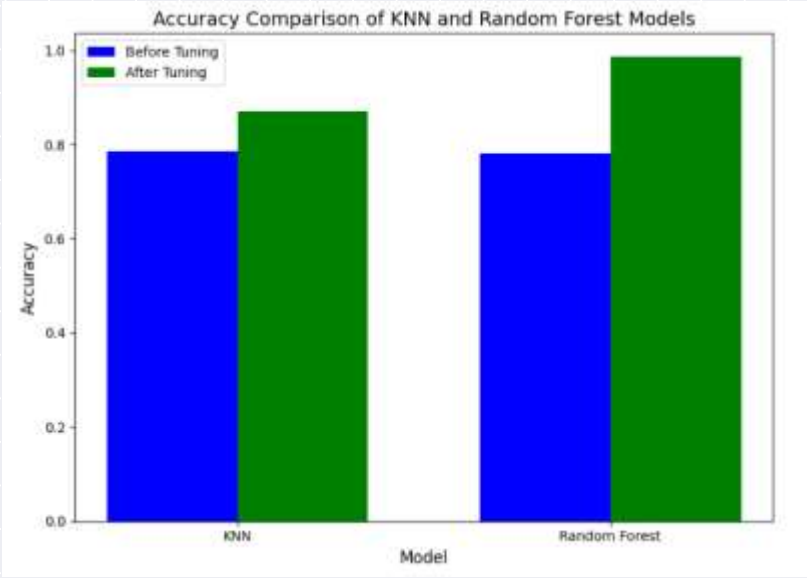
- K- Nearest Neighbour
- Random Forest



- Cross Validation
- Grid Search
- Oversampling (Imbalance dataset)

Predictive Modeling

Result for each model

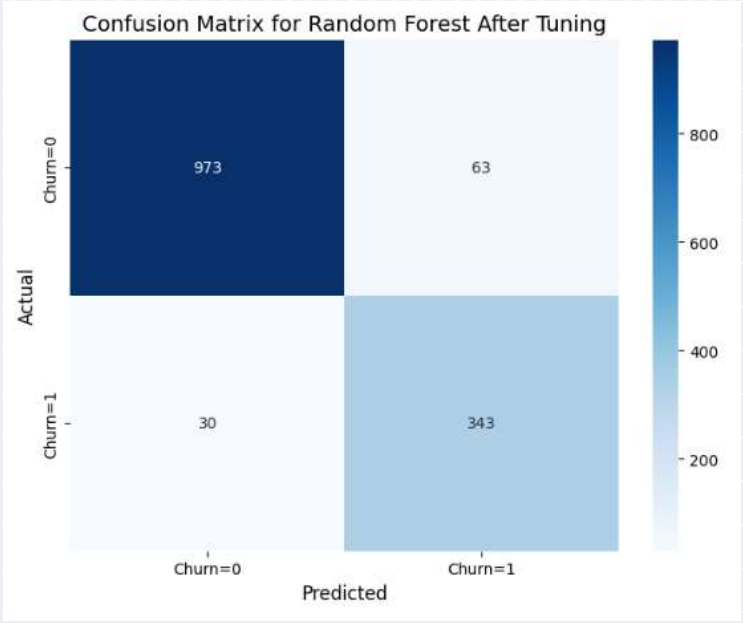



```
Classification Report Model Random Forest:
              precision    recall  f1-score   support

           0       0.97      0.94      0.95      1036
           1       0.84      0.92      0.88       373

    accuracy                           0.93      1409
   macro avg       0.91      0.93      0.92      1409
weighted avg       0.94      0.93      0.93      1409
```

- Evaluation for Random Forest with 93% Accuracy

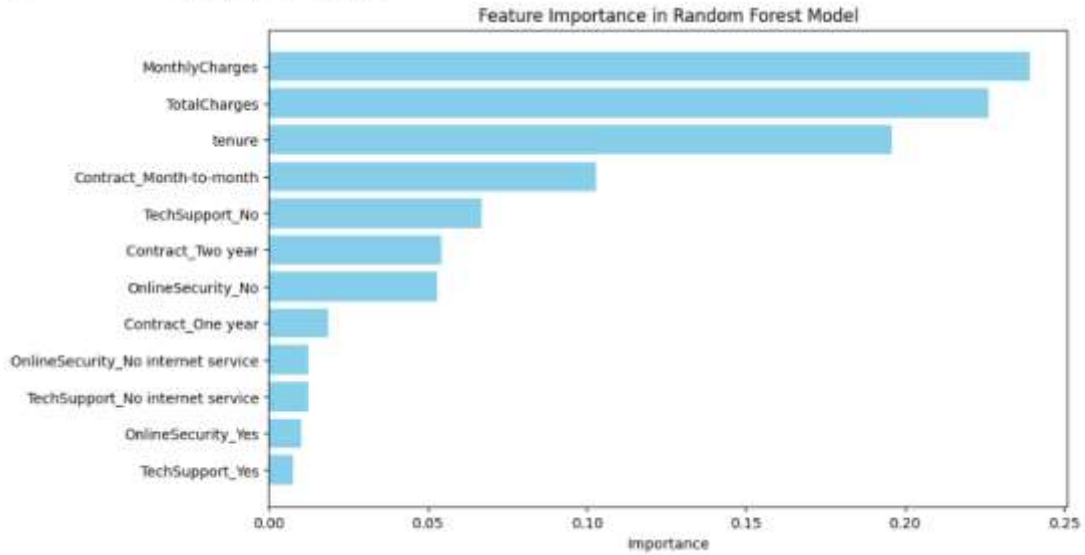


Predictive Modeling

- **Top Feature Impotance**

Feature Importances:

	Feature	Importance
1	MonthlyCharges	0.239185
2	TotalCharges	0.226259
0	tenure	0.196861
3	Contract_Month-to-month	0.102953
9	TechSupport_No	0.066975
5	Contract_Two year	0.054188
6	OnlineSecurity_No	0.053037
4	Contract_One year	0.018485
7	OnlineSecurity_No internet service	0.012571
10	TechSupport_No internet service	0.012327
8	OnlineSecurity_Yes	0.010221
11	TechSupport_Yes	0.007728



Explanation

- Very detail on feature importance compare with before because this dataset using One Hot Encoding.

Provide Solution

Recommendations for reducing churn

- ➔ **Focus on Customers with Month-to-Month Contracts:** Introduce Rewards Programme
- ➔ **Focus on High Total Charges Customers:** Cash back
- ➔ **Addressing Issues with OnlineSecurity and TechSupport:** 24/7 Technical support and online security
- ➔ **Utilizing ML Models for Churn Prediction:** Early Warning Sytem for Retentation

Project Timeline

Week 1 & 2

- Project Planning & Setup (2 Weeks)

Week 3 - 5

- Data Collection & Preparation (3 Weeks)

Week 6 - 9

- Modelig and Analysis (4 Weeks)

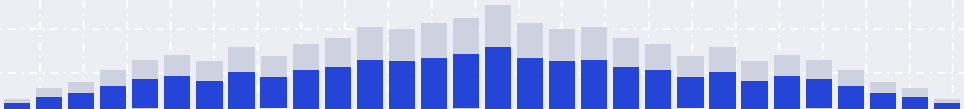
Week 10

- Final Review and Documentation (1 Weeks)

Project Budget

Bil	Cost Component	Cost (RM)	Explanation
1.	Salaries	15,000	Payment to team data experts for building a predictive model.
2.	Infrastructure	12,000	server (RM 10,000), and cloud storage(RM 2,000).
3.	Testing & Maintenance	4,000	Estimated cost for testing and updates (RM 1,000 per month for 4 months).
4.	Survey Costs	4,000	Includes survey design (RM 1,000), distribution (RM 500), incentives (RM 1,500), and data analysis.

Total Cost: RM 35,000



Conclusions

- In the customer churn analysis for the companies, we successfully developed a classification model with 93% accuracy.
- This model identifies customers at risk of churning and highlights key factors influencing their decisions, such as pricing, service quality, and customer support.
- Implementing these strategies will help proactively engage with at-risk customers, reduce churn rates, and improve overall customer satisfaction.
- Customer retention is vital for the growth and profitability of telecommunications companies like Celcom, etc.
- By leveraging data-driven insights and implementing targeted strategies, can strengthen its customer relationships and gain a competitive advantage in the market.

Thank You For Your Time !