Enhancing Customer Retention: Algorithms for Customer Churn Prediction

Anastasiya Jilani
School of Technology
New Zealand School of Education
Auckland, New Zealand
anastasiya.jilani@gmail.com

Abstract— **Customer churn is the exit of customers from a business and may have critical impact an organisation's growth and long-term profitability. The likelihood of churn can be predicted leveraging artificial intelligence algorithms, mitigating churn risk through customer retention strategies.**

Competitor advantage, customer dissatisfaction of product or service, pricing and marketing were identified as main reasons for churn. The major implications of customer churn on businesses are identified by literature as negative impact on profitability, growth and increased cost through of acquiring new customers. Data serves as inputs to churn prediction algorithms; however, sources of data are often underutilised by businesses and could help improve model performance and mitigate churn risk.

A literature review explored several high-performing models for customer churn prediction and their advantages and disadvantages. Algorithms aimed to predict customer churn as well as mitigate churn risk through behavioural analysis prior to churn, understanding underlying reasons for churn as well as churn within customer segments.

The aim of this paper was to explore, apply and compare models for capable of churn prediction. The dataset selected aimed to divert from typical industries and explore integration of methods. This study compared several models on a Travel dataset. Decision Tree and Random Forest models achieved highest accuracy of 90% and 87% respectively, performing higher than other models in line with literature findings. Ways to overcome limitations of static, single-source data and use real-time streaming and big data were explored as to provide a truer picture of modern business and help businesses understand reasons for retention. This could allow better, faster churn prediction models enabling faster intervention and mitigation in the future.

Keywords— Customer Churn, Predictive Analytics, Random Forest Classifier, Machine Learning

I. Introduction

Customers are the major source of profit and management of customer churn is vital to for company survival. Retaining customers is cheaper than acquiring new ones [4]. Churn implications for business include negative impact on growth and profitability, sales revenue, market share and brand image [1]. Several influences impact customer decision to churn, which are explored in this paper. In this study, algorithms used for churn prediction are explored. Predictive analysis using algorithms and input of data accumulated by firms can help gauge understanding for churn as well as timely prediction to enable proactive mitigation and retention strategies.

1.1. Problem Statement

To explore algorithms for churn prediction and integrations to improve predictive analysis.

1.2 Aims and Objectives

The aims and objectives of this study are to:

- Explore artificial algorithms for churn prediction.
- Explore limitations, advantages and disadvantages of methods.
- Explore integrations such as big data, time-variables to improve predictive analysis.
- Propose a method for churn prediction.

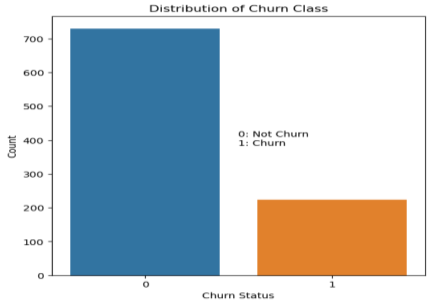


Figure 1 Exploring churn balance in dataset.

Limitations of this study focused best models identified in literature. Methods to improve models include leveraging big data sources, time variables and were limited to explanations of possibilities in previous studies. Historic travel data was used as it was not possible connect to cloud services for real-time streaming of data sources.

The remaining structure of this paper explores the problem statement, aims and objectives. Next, a literature review explores two themes 1. Algorithms and 2. Advantages and Disadvantages. The next section shares

Opinion on the subject matter, presenting gaps and methodologies.

II. LITERATURE REVIEW

Literature reviewed two main themes: algorithms for churn prediction and their limitations, advantages and disadvantages. The focus was algorithms and ways to optimise models that predict customer churn in presence of modern business implications. Machine learning, deep learning and hybrid methods for churn classification were examined.

Customer churn or attrition rate is the rate at which a customer discontinues doing business with a firm [5]. Customers may exit a business voluntarily or involuntarily for reasons such as switching to better competitor offerings, changes in product or service support, quality, price and overall customer dissatisfaction. Growth of technology and web enables competition through sale of goods on e-commerce and rise of online and cloud-based service platforms. Customers churn decision could be impacted by personal circumstances, disposable income, economic and political changes[4].

Social media also influences product and service trends, providing customers a platform for expressing praise and dissatisfaction as well as scrutinise business actions. Businesses may lose customers due to negative reviews of product or service.

Churn implications for business include negative impact on growth and profitability, sales revenue, market share and brand image [1]. Customers are the major source of profit and management of customer churn is vital to for company survival. Retaining customers is cheaper than acquiring new ones [4].

Literature mentions although companies accumulate big data - of huge volume, velocity and variety - from various sources, it is underutilised for churn mitigation. Knowledge from customer data can be extracted to know reasons for churn and retain customers[4]. Customer dissatisfaction is a key reason for churn however business-specific reasons could also exist. Prediction models can also help alert businesses early to provide better service [5]. Utilising methods to understand reasons for customer behaviour, churn risk and prediction allow intervention through customer relationship management (CRM), marketing, loyalty and retention strategies to help retain profits[1].

Multi-dimensional data sources and time variables as inputs to algorithms can provide a clearer picture of churn. B2C (business to customer) interactions capture multi-dimensional data. Big data significantly optimised progression of estimating customer churn[3]. Big data tools integrated with cloud technology can aid extraction, transformation and loading of data along the pipeline for use in predictive analysis.

The literature found classification algorithms capable of predicting customer churn, however focused on historic data used mostly contract-bound service businesses. This study explores and compares algorithms and explores ways to integrate methods that may improve customer churn prediction and analysis.

Several studies leveraged Machine Learning (ML) binary classification algorithms. A study explored their use for churn prediction in telecommunications to improve business intelligence and customer relationship management[7]. It stated, other studies lacked consideration for real-world implications and missing data causing poor or biased prediction results. XBoost algorithm outperformed Support Vector Machines, K-Nearest Neighbour (K-NN), Decision Trees and Random Forest Classifier. Xboost achieved 96% accuracy with 96% precision which is the ability to correct identify customers that churned. A Confusion Matrix was used to evaluate false positive and false negatives. Optimisation of models including refining accuracy through hyperparameter tuning using Random-Search or Grid-Search techniques to find best hyperparameter inputs.

Integration of Neural Networks (NNs) with ML models could increase their capabilities and integration of data sources [5]. NNs have also been used in a number of industries to predict customer churn. A study used a weights and structure determination (WASD) based NN for customer churn prediction in telco with 82% accuracy, outperforming the other models in the test which were fine tree (77%), kernel naïve Bayes (81%) and fine K-nearest neighbours (71%)[6].

Hybrid models combine methods such as Yuchen Jiang's [1] model using clustering and classification on banking dataset. Significance features within clusters was analysed for churn reasons. The 'Elbow method' was used to find 4 optimal clusters, Group 2 having a significant gap between churn (114,441) and non-churn (300) customers. Feature Importance Ranking allows features to be ranked based on impact on target variable. Support Calls, Total Spend and Age as top three variables impacting churn in Group 2. The likelihood of customers churning in category could be predicted and targeted retention taking place. Decision tree achieved 99% with Logistic Regression achieving 90% accuracy. Hybrid models require consideration of limitations of each model and accuracy for the scenario.

Another study applied a hybrid model to customer credit data for early churn, alerting companies to improve service levels in banking[12]. The two-fold approach analysing customer behaviour then building models. RF model achieved 90% accuracy. Irrelevant features were removed before training.

Xiahao and Harada [9] took a hybrid approach for B2C e-commerce emphasising the rising trend of e-commerce and online shopping. SVM and k-Means Clustering hybrid models were used. SVM predictions improved after clustering proving necessity of k-means clustering.

Mena et. al [5] utilised a hybrid approach using Deep Neural Network based predictions with conventional ML models. It was a simple method to extend ML to improve churn accuracy by incorporating time-varying data. RFM (Recency, Frequency and Monetary Value) data was integrated to evaluate customer risk to analyse customer behaviour over time.

Wu et. al [10] integrated a customer analytics framework with churn prediction, factor analysis, segmentation and behavioural analysis to provide a complete picture of churn analysis in telco. Overall, 6 classifiers were compared for 1. Predicting churn status. Then Bayesian Logistic was used for 2. Factor Analysis and k-means clustering for 3. Segmentation. The rationale is first accurately predicting churn, then understand reasons for churn then target retention strategies based on segmentation. Random Forest outperformed on 2 different datasets with 93.6% and 63.09% accuracy. F1 score was considered for accuracy with imbalanced dataset.

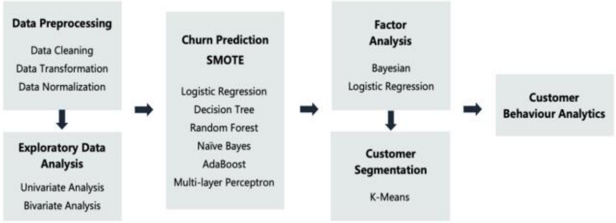


Figure 2 Integrated Telco Customer Analytics Framework by Wu et.al (2021).

Limitations, advantages and disadvantages of algorithms were considered. Imbalance of churn verses non-churn ratio needed consideration, contributing to poor model performance and bias, for example 84% churn and 16% non-churn in one study[7]. Standardisation and normalisation and Synthetic Minority Oversampling Technique (SMOTE) can be applied to deal with imbalanced datasets [10].

Feature ranking methods and irrelevant variables can be removed and be compared to industry importance [12] [1]. With hybrid models, consideration of model's principles was important to selection as it impacts accuracy and interpretability[1].

A major limitation is obtaining actual customer churn data due to company privacy. Most studies utilised historic data either virtual or real. Importance of real-world data, and missing data causing biased results and poor model performance was mentioned [7].

Another issue was focus on contract-bound customers e.g. in telecom and banking for predicting churn. Non-contract customers such as online shoppers can be more difficult to predict however accumulates big data through interaction[9]. Integration of multi-dimensional data can provide insights into customer behaviour and likelihood of churn [5]. Data sources such as abnormal website usage, subscription rate, interaction behaviour and social media were suggested.

Exclusion of longitudinal behaviour and longitudinal timeliness of customers was criticised [12]. Extracting customer activity data allowed modelling of behaviour and churn. Another study mentioned time importance to detect of behavioural trends churn prediction as customers do not go from loyal to ex-customers overnight [1]. Time was a crucial churn predictor for example in banks, lenders and service providers to assess behavioural trends prior to churn so analysis of daily, weekly, monthly and not just quarterly behaviour can be considered. Web interactions, subscriptions and social media with hybrid algorithms for churn prediction requiring integration of multi-dimensional data and customer behaviour over time [5][9]. CRM data was to be incorporated to obtain more complete reference of churn behaviour and applying retention.

Each algorithm presents its own advantages and disadvantages due to their characteristics and may be limited by use case, data type and size therefore model selection should incorporate all the above.

Decision Trees handle multi-categorisation problems, high dimensionality and large-scale datasets. They can handle missing values and robust to outliers [1]. However, they can be limited where complex, non-linear relationships between features exist but could be improved with linear features and hyperparameter pruning [7]. They are prone to overfitting for complex datasets and sensitive to small changes[1].

Logistic Regression enables simple, easy implementation and interpretation. It can also handle binary and multiple classification and is robust to outliers and noise. Its limitations include poor performance with non-linear problems and feature engineering affects input[1].

Random Forest has the best metrics in many studies [2][12][10]. Advantages include parallel scaling and coping with different variables and accuracy due to number of decision trees in the process. It has inherent feature importance contributing to target variable and avoids overfitting and bias by averaging predictions[12]. Limitations include less interpretability than other models, computational complexity and requiring tuning of hyperparameters[11].

Hybrid models using ML algorithms, clustering methods and even NNs can be useful if one method can overcome limitations of another. Principles of individual methods and dataset need to be considered.

III. Opinion

Based on the literature findings, businesses are implicated by customer churn and predictive enables can help businesses to save costs and maintain profits and growth. Churn prediction allows analysis of customer behaviour allowing for timely strategic decisions, customer interaction and retention planning.

Churn prediction for non-contract and e-commerce were more challenging as churn status is only proven by account deletion, non-payment or lack of contract renewal. Measures to retain customers include loyalty point systems, targeted advertising, discount coupons and subscriptions which may not be cost-effective for example, a customer churns immediately after receiving a discount or only purchasing when loyalty points are redeemable. Key reasons for churn included dissatisfaction, high costs, lack of support. Incidental churns factors include customer changes in finances or location. Deliberate churn includes factors include price, changes in tech, quality service and psychological reasons [4].

Therefore, analysis of the wider customer is necessary from multiple data sources. Static and stimulated data does not incorporate factors impacting churn including trends, social media semantics, economic, socio-political changes. Lack of real-world implications could lead to missing data and biased models[7].

Leveraging real-world customer data from multiple sources and not just personal data, as well as incorporation of longitudinal and latitudinal customer behaviour provides a more completed picture of customer[9]. Insights from other studies show methods to fill gaps for churn prediction by integrating CRM platform[5] multiple data sources and time frames [9] and integrated frameworks[10][2] could provide a more complete picture of churn prediction, reasons and ability to perform retention plans. Churn doesn't occur overnight[1] so timely analysis provides earlier detection and prevention of churn.

Businesses need to understand reasons for churn with multi-dimensional data providing deeper insights. While dimensionality may be introduced to data, feature ranking either inherent in models or through feature scoring methods can help determine greatest impact on customer churn. Businesses could make data-driven decisions to detect churn, allowing timely mitigation of churn risk, and intervention through CRM, retention strategies and methods like Life-Time Value and Cost of Intervention.

A solution could incorporate the above as a multi-analysis framework. In this experiment prediction models are compared for churn prediction accuracy. We explored integrations and later models could include incorporating big data and real-time analysis.

Data sources would go beyond personal data and service or site use but also integrate social media semantic text analysis, feedback channels, reviews and web interactions. Further integration could include breaking down data via customer segmentation of customer types, value or churn risk potential. Timeframes can be incorporated based on business significance – daily, weekly, quarterly and even daily behaviour.

Time segmentation, customer segmentation, multiple big data sources could be utilised with real-time data stream – and visualised as real-time metrics displayed on dashboard that alert business managers and could include automated customer retention plans or traditional methods such as email and phone to communicate customer status.

Based on the literature review machine learning models the such as Random Forest, Decision Trees and Logistic Regression have high accuracy and are simple, easily implemented non-hybrid models [12][2]. Based on this the RF model will be implemented its performance compared with other models.

Methodology

This section outlines the steps to create a customer churn prediction model using Random Forest Classifier algorithm. First, the model is built following a general machine learning framework, then compared to other models using evaluation metrics. The complete source-code for this project is provided on GitHub link[13].

Dataset

Travel and Tours customer database was used[8]. This dataset was selected firstly to make it simpler to focus on various models' performance on the same data without too many dimensions. Secondly, it was difficult to obtain real customer data for ML or datasets outside of banking and telco.

Tools

Python programming language and libraries (e.g. Pandas, Matplot, StandardScaler) were imported for all data processing, manipulation and visualisation in the Jupyter Notebook coding platform.

Evaluation Methods

ML Classification models were evaluated using classification metrics used in literature including Accuracy, Precision, F1 Score, Recall and Confusion Matrix.

Implementation of churn prediction model

1) Select Dataset and Tools

The first step was to select a dataset, Travel and Tours Customer. This is based on historical data with features as below.

FEATURES (X)	TARGET (Y)
	Target (Customer Churn)
AGE	
FREQUENT FLYER	
ANNUAL INCOME LEVEL	
NUMBER SERVICES OPTED	
SOCIAL-MEDIA SYNC	
BOOKED HOTEL OR NOT	

2) *Data Storage*

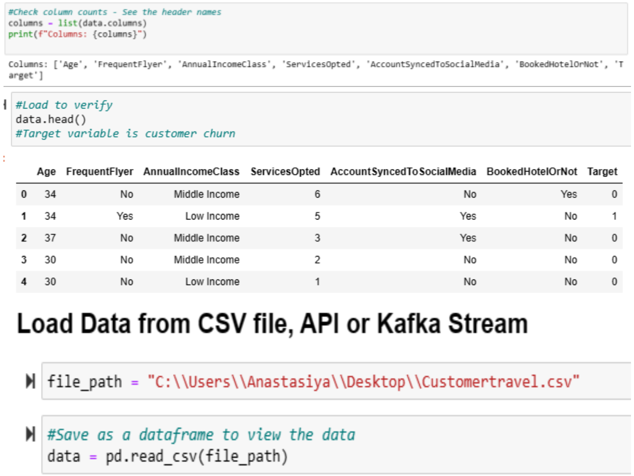
Identify storage source could be CSV, Parquet, JSON file from device or cloud storage, Kafka stream or database.

3) *Data Extraction or Loading*

Data was loaded by identifying file storage path the reading as Pandas DataFrame using pd.read_csv function.

4) *Verify Loading of Data*

Data head and tail is loaded. Verify rows and header names, adjusting names as necessary.



```
#Check column counts - See the header names
columns = list(data.columns)
print(f"Columns: {columns}")
```

```
Columns: ['Age', 'FrequentFlyer', 'AnnualIncomeClass', 'ServicesOpted', 'AccountSyncedToSocialMedia', 'BookedHotelOrNot', 'T
arget']
```

```
#Load to verify
data.head()
#Target variable is customer churn
```

	Age	FrequentFlyer	AnnualIncomeClass	ServicesOpted	AccountSyncedToSocialMedia	BookedHotelOrNot	Target
0	34	No	Middle Income	6	No	Yes	0
1	34	Yes	Low Income	5	Yes	No	1
2	37	No	Middle Income	3	Yes	No	0
3	30	No	Middle Income	2	No	No	0
4	30	No	Low Income	1	No	No	0

Load Data from CSV file, API or Kafka Stream

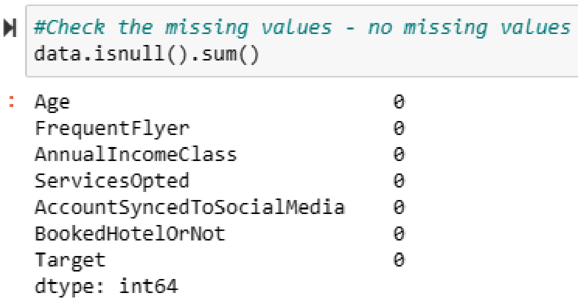
```
file_path = "C:\\Users\\Anastasiya\\Desktop\\Customertravel.csv"
```

```
#Save as a dataframe to view the data
data = pd.read_csv(file_path)
```

5) *Data-preprocessing*

Data pre-processing is to clean the data to make it acceptable for input. Data is cleaned for noise, filtering outliers, duplicates, erroneous data, missing values, datatypes.

Check Missing Values



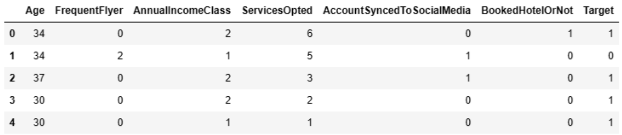
```
#Check the missing values - no missing values
data.isnull().sum()
```

```
Age                           0
FrequentFlyer                 0
AnnualIncomeClass             0
ServicesOpted                 0
AccountSyncedToSocialMedia    0
BookedHotelOrNot              0
Target                        0
dtype: int64
```

6) *Encoding Categories*

Categorical variables are encoded to numerical to allow algorithm to process it. This step preserves categorical information, avoids bias and improves performance.

One-Hot encoding to assign binary codes to non-ordinal data. Label encoding numbers of categories for ordinal ranking.



	Age	FrequentFlyer	AnnualIncomeClass	ServicesOpted	AccountSyncedToSocialMedia	BookedHotelOrNot	Target
0	34	0	2	6	0	1	1
1	34	2	1	5	1	0	0
2	37	0	2	3	1	0	1
3	30	0	2	2	0	0	1
4	30	0	1	1	0	0	1

7) *Visualisaton and Exploratory Analysis*

Visualising data is important to examine relationship of Target with Features, check the distribution, balance and linearity of data.

a) Distribution of Churn vs Non-Churn

The bar plot shows count of churned customers.

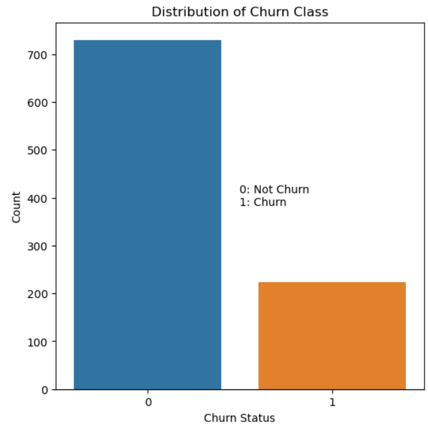
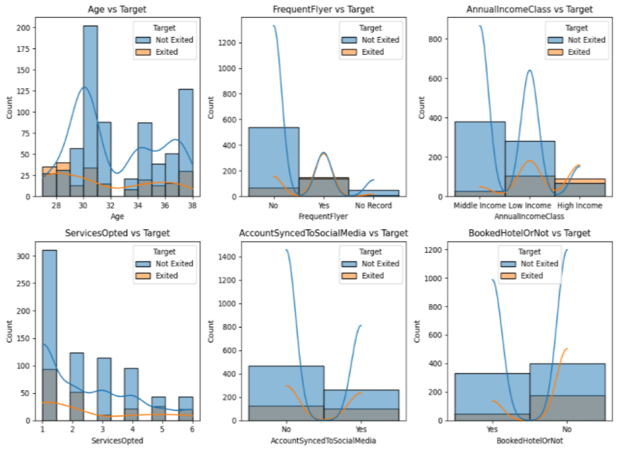


Figure 3 Churn v Non-Churn Customers

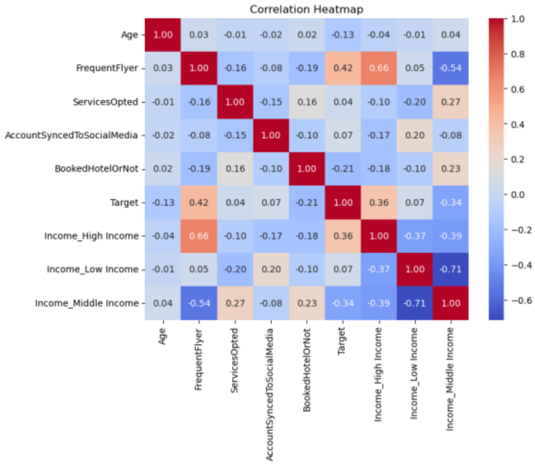
b) Distribution of Churn within Features

The proportion of the Target variable within each feature is explored including their distribution. Feature engineering methods can determine impact of each feature.



c) Correlation Heatmap

Correlation between variables can indicate positive and negative relationships between variables but does not always imply causation.

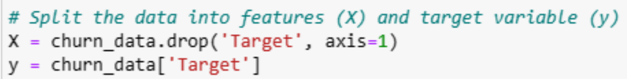


8) *Feature Engineering.*

For datasets with many dimensions feature engineering is important to avoid curse of dimensionality, selecting most impactful features. However, models such as RF have inherent scoring which can be view after. Due to small number of features all 6 were included.

9) *Separate Target and Features*

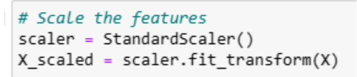
The models will explore impact of variables on customer churn. Statistically this is the impact of (X/ independent features) on (y/dependent target), so the dataset is split.



```
# Split the data into features (X) and target variable (y)
X = churn_data.drop('Target', axis=1)
y = churn_data['Target']
```

10) *Feature scaling*

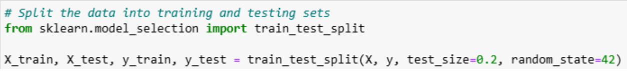
The features are normalised or standardised using StandardScaler which ensures same scale and fairness, avoiding dominance, and improve interpretability by making weights more meaningful.



```
# Scale the features
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
```

11) *Split into Training and Testing Sets*

Using train_test_split from Scikit-Learn library and a ratio of 80/20 the dataset is split. 80% is used for training and 20% for testing models capabilities on unseen data.

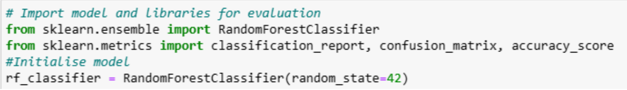


```
# Split the data into training and testing sets
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

12) *Model Initalisation*

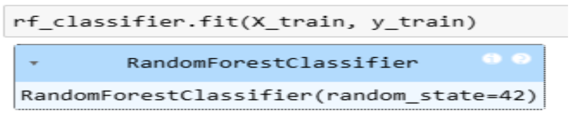
The model(s) selected are imported from Scikit-Learn and initialised including Random Forest Classifier which is described here.



```
# Import model and libraries for evaluation
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score
#Initialise model
rf_classifier = RandomForestClassifier(random_state=42)
```

13) *Model Training – Fitting*

The model is fitted with the training data in to learn the patterns within variables to predict for y Customer Churn.

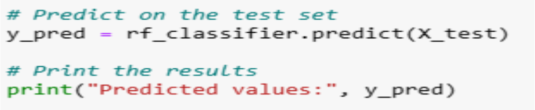


```
rf_classifier.fit(X_train, y_train)
```

```
RandomForestClassifier(random_state=42)
```

14) *Model Testing - predictions*

The RF model is run, and its predictive capabilities are tested on the unseen data. The results of the prediction are compared with actual dataset values.



```
# Predict on the test set
y_pred = rf_classifier.predict(X_test)

# Print the results
print("Predicted values:", y_pred)
```

Target	Actual	Predicted
1 (Churn)	153	159
0 (Non-Churn)	38	32

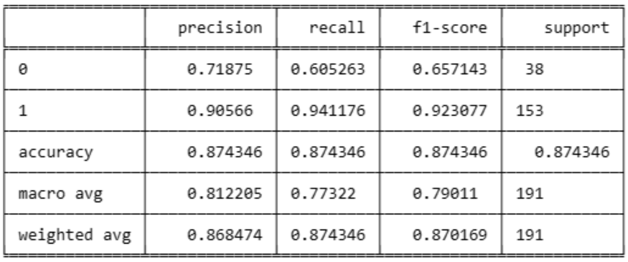
Figure 4 RF Churn Prediction Results

15) *Evaluation Metrics*

To assess the model’s performance, evaluation metrics are used including Accuracy, Precision, F1 Score, and Confusion Matrix are used.

a) *Classification Report*

An overall accuracy score of 0.87% which is comparable to RF models in the literature.

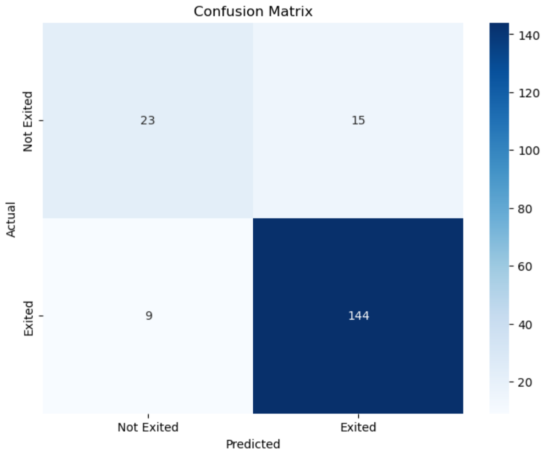


	precision	recall	f1-score	support
0	0.71875	0.605263	0.657143	38
1	0.90566	0.941176	0.923077	153
accuracy	0.874346	0.874346	0.874346	0.874346
macro avg	0.812205	0.77322	0.79011	191
weighted avg	0.868474	0.874346	0.870169	191

Figure 5 Classification Report. 1=Churned, 0 = Non-Churn

b) *Confusion Matrix*

The confusion matrix helps understand performance of the model based on True/False Positive and Negative classifications. A high number (144) of True Positives means it correctly identified churn where customer churned.



16) *Intepretation of Results*

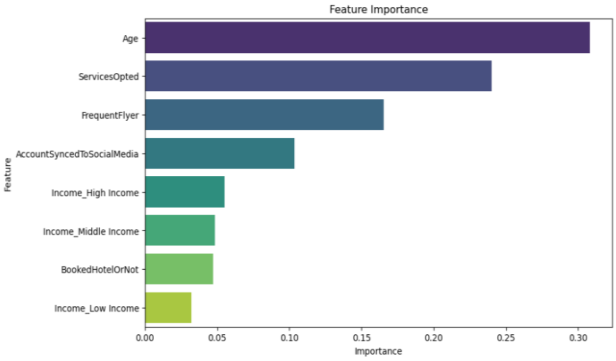
The model has a high accuracy of 87.43% overall comparable to RF models in the literature. The precision for Class 1(churned) (90.57%) means the model performs well predicting churn and with a high recall rate (94.12%) suggesting successful identification. The F1 score is also high demonstrating robustness for predicting Class 1. High true positives (144) mean the primary objective of identifying churn is met and retention strategies can be undertaken.

However, Precision (0.71%) and Recall (0.66%) for Class 0(Non-Churned) is lower meaning some false negatives may occur. Overall, the confusion matrix identifies (9) false negatives who actually churned and 15 false positives who did not churn. This could mean lost opportunities for retention or wasted resources on retention.

17) *Feature Ranking*

RF Classifier Feature Importance was used to assess features impacting churn the most. These were Age, Services Opted and Frequent Flyer, indicating key reasons for churn for business to focus on.

a) Feature scoring based on Random Forest



18) *Next Steps – Optimisation, Storage and Use*

Model re-training on new data can make it more accurate. The model can be saved as package and applied at the data analysis pipeline to incoming data from various sources. The results can be stored via database or cloud. Integration of real-time data and dashboards can help data-driven decision making.

IV. Comparison of Models

The process was repeated for various models as summarised below with Decision Tree obtaining the highest accuracy and precision of (90%) followed by Random Forest (87%). This could be because Decision Trees can deal with complex features and maintain interpretability. Logistic Regression and K-Nearest Neighbour also performed well at 86% overall accuracy. The SVM model had weaker precision. Each models performance varies based on dataset, features and tuning and its ability to deal with noise, overfitting and non-linearity.

Table 1 Comparison of Models

Model	Precision	Recall	F1-Score	Accuracy
Random Forest	0.868474	0.874346	0.870169	0.874346
Decision Tree	0.898041	0.900524	0.898962	0.900524
Logistic Regression	0.855732	0.863874	0.857601	0.863874
SVM	0.641676	0.801047	0.712559	0.801047
K-Nearest Neighbour	0.861330	0.863874	0.862477	0.863874

Furthermore, feature importance score may change e.g. Logistic Regressor Model emphasised 'BookedHotelOrNot', 'IncomeMiddle' and 'AccountSyncedToSocialMedia' as higher importance compared to RF and Decision Tree.

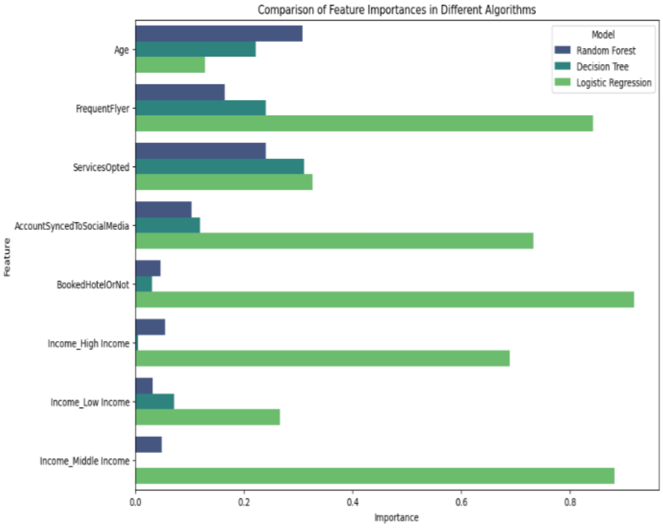


Figure 6 Comparison of Feature Scores based on models.

Limitations and Improvements

In this study, the accuracy of models on churn prediction were explored and methods to improve overall models were identified. The first gap is data quality. More sources of data (social media, web interaction, surveys), larger datasets and real-time data can be utilised. A data ingestion pipeline incorporating cloud and distributed storage systems enables streaming of data, pre-processing and deployment of a model like the ones examined. Seconding, feature scoring differed and metrics differed between models though models can be optimised through hyperparameter tuning and advanced feature scoring methods. Neural networks can understand greater complexity in features but might work better for larger datasets. A third major limitation was access to customer databases due to ethics, although methods like social media analysis can be helpful to gauge customer semantic. Another limitation is interpretation for decision makers and timeliness since customers do not churn overnight. Real-time dashboards and integration with business intelligence tools could aid with visualisation. Segmenting results based on time-risk and churn-risk of customers to use pro-active retention measures.

V. Conclusion

Customer churn prediction is critical for businesses to grow, retain profits and save costs. Customers have various reasons for churn. Artificial intelligence algorithms can help predict churn, understand reasons for churn and be integrated and utilised for further marketing metrics, retention and customer relationship management.

The quality and performance of models rely on quality data, with wider sources of data providing a clearer picture. Timeliness in prediction is essential for timely intervention. Machine learning and deep learning models are useful methods to predict churn. The best performing models for customer churn prediction in this study were Decision Tree (90%) and Random Forest (87.43%) with comparable levels of accuracy to literature. The high precision and true positives indicate achieve the intention of predicting possible churn on the Travel dataset.

The high accuracy and precision of models, integrated with further data and visualisations could help businesses make data-driven decisions. Timely risk assessment, prediction as well as target retention can help save costs from acquiring new customers, retain brand loyalty, improve brand image and overall growth and profitability.

Future works would improve the model through optimisations such as hyperparameter tuning, SMOTE, using diverse data sources such as social media semantics and explore scalability of models. Clustering could be used for customer churn segmentation to improve CRM and retention strategies. Use of real-time data ingestion pipeline and big data tools in conjunction with live dashboards could provide real-time analysis for timely interaction and retention which future works could examine.

References

[1] Cheng, L. C., Wu, C.-C., & Chen, C.-Y. (2019). Behavior Analysis of Customer Churn for a Customer Relationship System: An Empirical Case Study. *Journal of Global Information Management*, *27*, 111–127. https://doi.org/10.4018/JGIM.2019010106

[2] Jiang, Y. (2024). Customer Churn Analysis Prediction Based on Cluster Analysis and Machine Learning Algorithms. *Advances in Economics, Management and Political Sciences*, *77*, 192–198. https://doi.org/10.54254/2754-1169/77/20241662. Retrieved 01 May 2024.

[3] Joolfoo, M. (2020). Customer Churn Prediction in Telecom Using Machine Learning in Big Data Platform. Retrieved 05 May 2024.

[4] Mathai, P. (2020). Customer Churn Prediction: A Survey. *International Journal of Advanced Research in Computer Science* Volume 8, No. 5, May – June 2017. www.ijarcs.info. Retrieved 07 May 2024.

[5] Mena, G., Coussement, K., De Bock, K., De Caigny, A., & Lessmann, S. (2023). Exploiting time-varying RFM measures for customer churn prediction with deep neural networks. *Annals of Operations Research*, 1–23. https://doi.org/10.1007/s10479-023-05259-9

[6] Mourtas, S. (2024). Customer churn classification through a weights and structure determination neural network. *ITM Web of Conferences*, *59*. https://doi.org/10.1051/itmconf/20245901004. Retrieved 01 May 2024.

[7] Sam, G., Asuquo, P., & Stephen, B. (2024). Customer Churn Prediction using Machine Learning Models. *Journal of Engineering Research and Reports*, *26*, 181–193. https://doi.org/10.9734/jerr/2024/v26i21081. Retrieved 01 May 2024.

[8] Tour & Travels Customer Churn Prediction. Predict Tour & Travels Customer Churn. Kaggle. Tour & Travels Customer Churn (kaggle.com). Retrieved 13th March 2024.

[9] Xiahou, X., & Harada, Y. (2022). B2C E-Commerce Customer Churn Prediction Based on K-Means and SVM. *Journal of Theoretical and Applied Electronic Commerce Research*, *17*(2), Article 2. https://doi.org/10.3390/jtaer17020024. Retrieved 01 May 2024.

[10] Wu, S., Yau, W.-C., Ong, T.-S., & Chong, S.-C. (2021). Integrated Churn Prediction and Customer Segmentation Framework for Telco Business. *IEEE Access*, *9*, 62118–62136. https://doi.org/10.1109/ACCESS.2021.3073776

[11] Zandi, S. (2024). Week Four: Business Decision Making Approaches with Big Data. GDDA709 Big Data Analytics. New Zealand School of Education Course Materials. Week 4 : GDDA709 Big Data Analytics (instructure.com)

[12] Zhao, S. (2023). Customer Churn Prediction Based on the Decision Tree and Random Forest Model. *BCP Business & Management*, *44*, 339–344. https://doi.org/10.54691/bcpbm.v44i.4840. Retrieved 01 May 2024.

[13] Jilani, A (2024). GDDA709 BigDataAnalytics. Travel Customer Churn Prediction Source Code. anastasiya-j/GDDA709-BigDataAnalytics: Python Code and Dataset Used for Customer Churn Prediction. (github.com)