Machine Learning and Predictive Analysis

Dhanushka Gunasinghe

23052728

Abstract

This study and analysis performed to provide effective machine learning based predictive analysis and decision-making model to provide effective solution for churn management of Sri Lanka Telecom. In additionally identified the features or attributes that highly co-related with customer churn, and provide opportunity to address the findings and issues.

Source dataset and the jupyter notebook for running the analysis are available here.

https://gitlab.uwe.ac.uk/d2-gunasinghem/sri-lanka-telecom_churn-prediction/-/tree/main

Introduction

Sri Lanka Telecom is the National Telecommunication provider in Sri Lanka played a crucial role in the development and evaluation of the Telco sector. Sri Lanka Telecom was established in 1858 during the British colonial era, since from the establishment company has maintained its leadership and guided to the Telecommunication sector in the country. After year 2000 Sri Lankan government liberalize the telecommunication market for private and global players. After the market revolution global players like Bharati Airtel, Dialog Axiata established the country making telco landscape more competitive and dynamic. With the higher level of competition, Annual revenue of Sri Lanka Telecom drastically decreased. The survey conducted by an independent agency found that churn existing customers and move into competitors is the main reason behind the decrease. This paper describes [1] importance of the churn protection management for any kind of business domain in competitive landscape.

This study and analysis are crucial for the Management of Sri Lanka Telecom to **determine the key factors that contributing to the churn**. **Pro-actively detect the potential churn customers** earlier and safeguarding customers as well as Sri Lanka Telecom's revenue and establishment.

Machine Learning for Churn Prediction

Machine Learning enables businesses and management to make quick, effective, and accurate decisions based on proven mathematical algorithms.[2] [3][4]. Especially in the context of Telecommunication under that Is more important and valuable in highly competitive environments.[5]

Scope

- Develop and evaluate machine learning model based on different classification algorithms and provide effective predictive analysis solutions for Sri Lanka Telecom to enable early detection of potential retention customers.
- Perform feature evaluation and analysis, to determine highly co-related features.

Dataset

Sri Lanka Telecom CRM Dataset (2023), Under fully approval of Manager-Data Analytics. This Dataset contain 100000 records with 100 feature attributes distributed entire regions in Sri Lanka.

Exploratory Data Analysis

The Dataset contains 100000 records with 100 feature attributes.

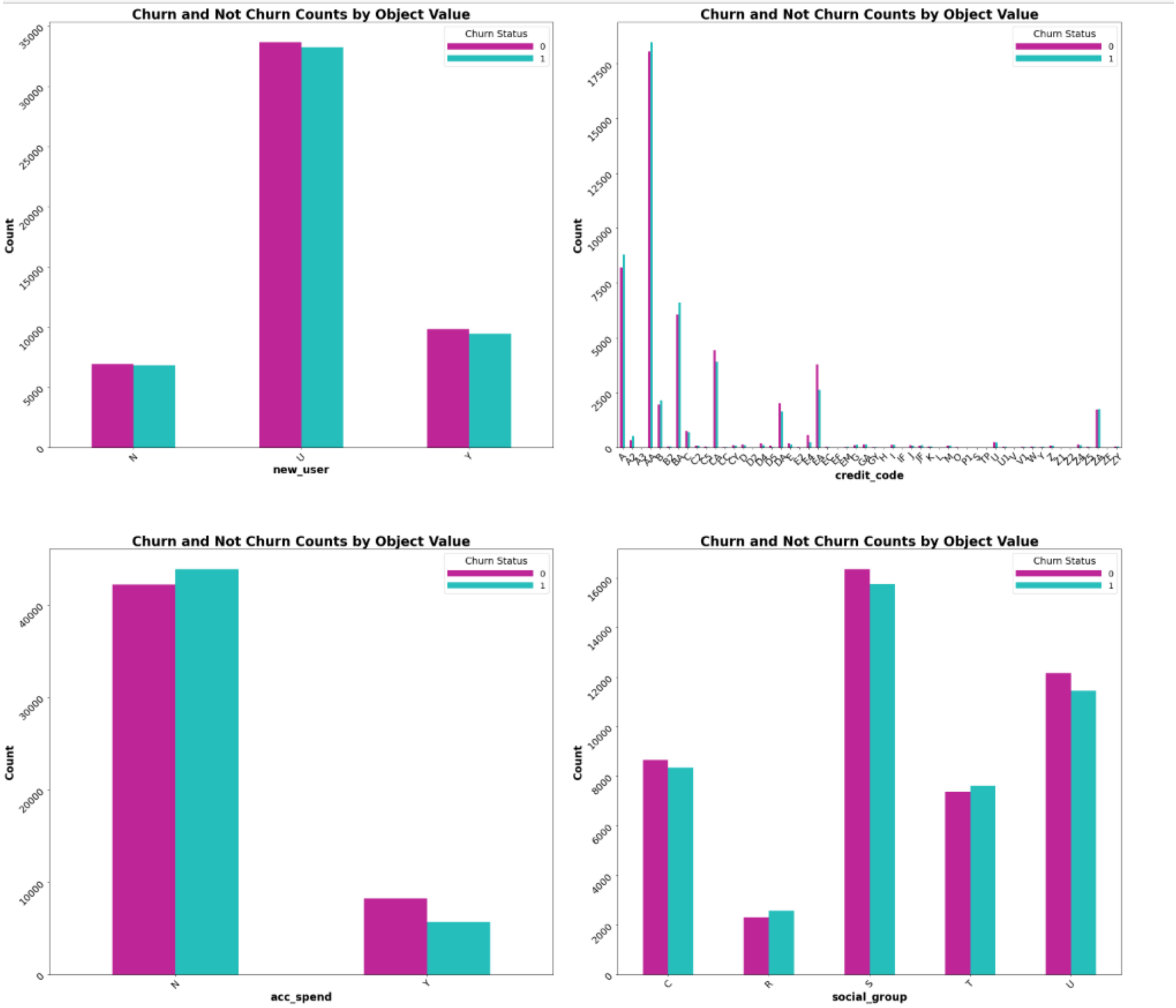
21- Categorical feature attributes

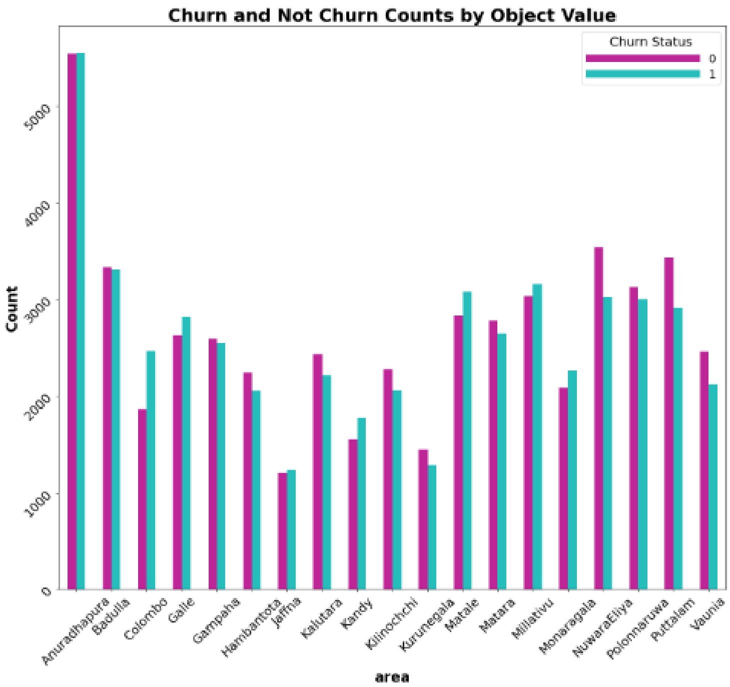
79- Continues feature attributes

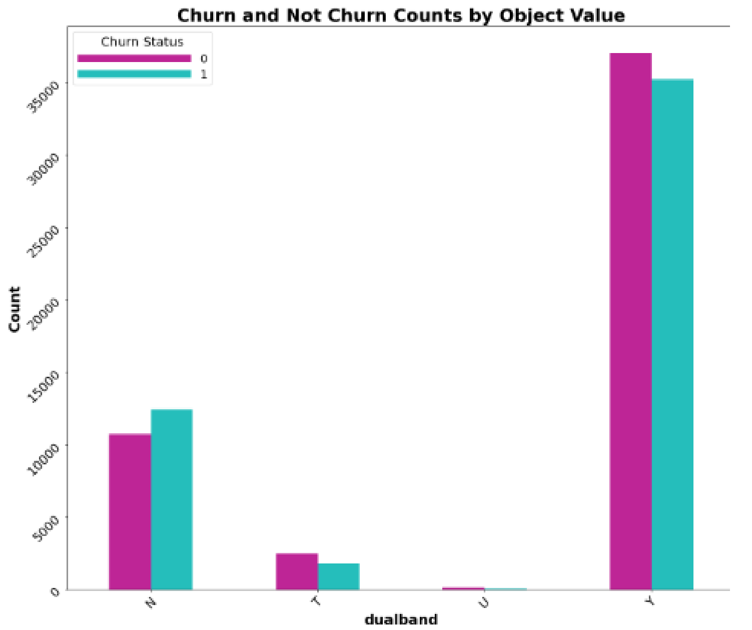
	new_user	credit_code	acc_spend	social_group	area	dualband	refurb_new	hnd_webcap	ownrent	dwlltype	...	infobase
0	U	A	N	S	Colombo	Y	N	WCMB	O	S	...	M
1	N	EA	N	U	Gampaha	N	N	WC	NaN	S	...	M
2	Y	C	N	S	Kalutara	N	N	NaN	O	S	...	M
3	Y	B	N	T	Gampaha	N	N	NaN	NaN	M	...	M
4	Y	A	N	U	Galle	Y	N	WCMB	R	M	...	M
...	...	...	...	...	...	...	...	...	...	...	...	...
99995	U	B	N	U	Badulla	N	N	WC	O	S	...	M
99996	U	CY	Y	S	Badulla	N	N	WC	O	S	...	M
99997	U	DA	N	U	Millativu	Y	N	WCMB	NaN	NaN	...	M
99998	U	EA	N	U	Millativu	Y	N	WCMB	NaN	NaN	...	NaN
99999	U	B	N	S	Badulla	Y	N	WCMB	NaN	S	...	M

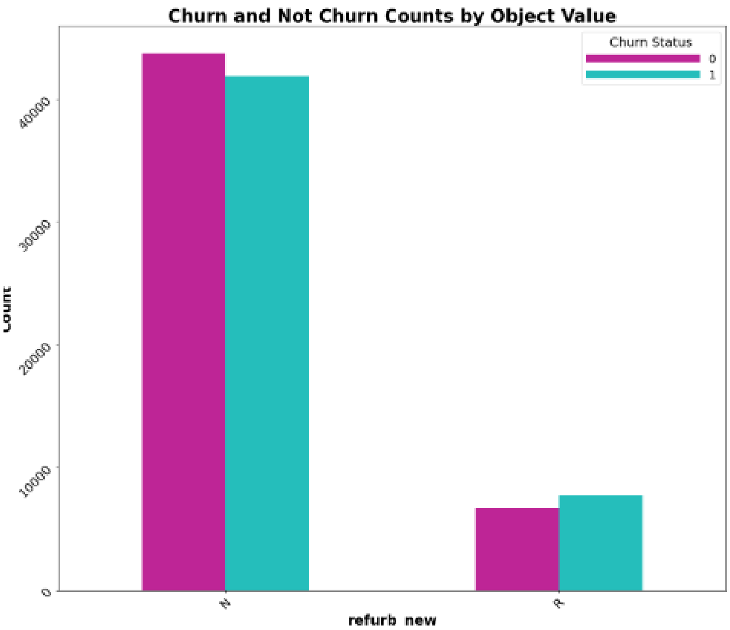
100000 rows × 21 columns

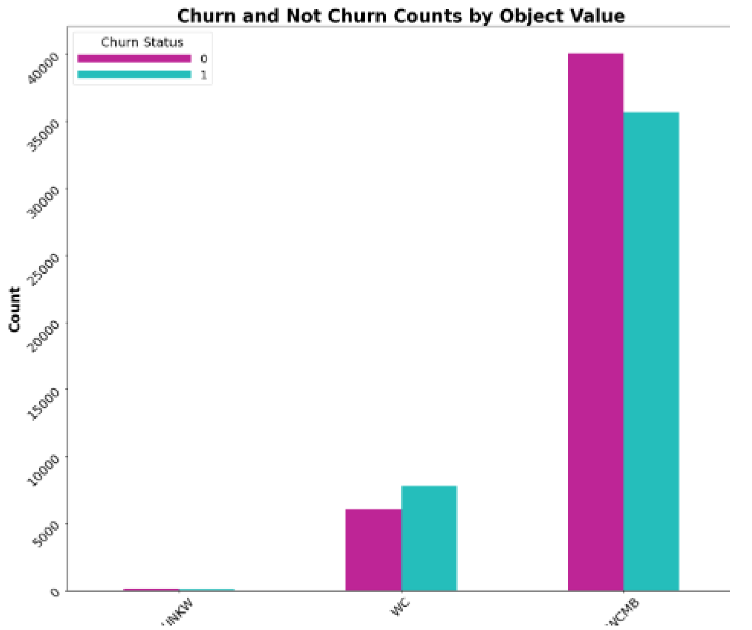
My First observation categorical variable behavior with the churn status.

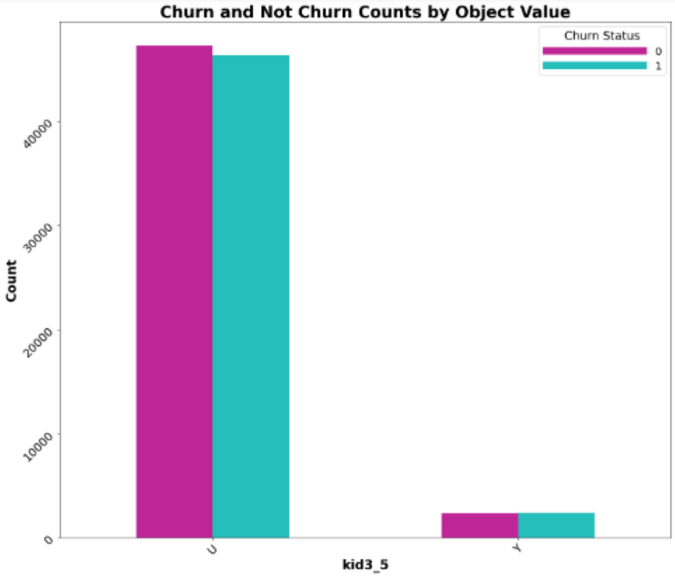


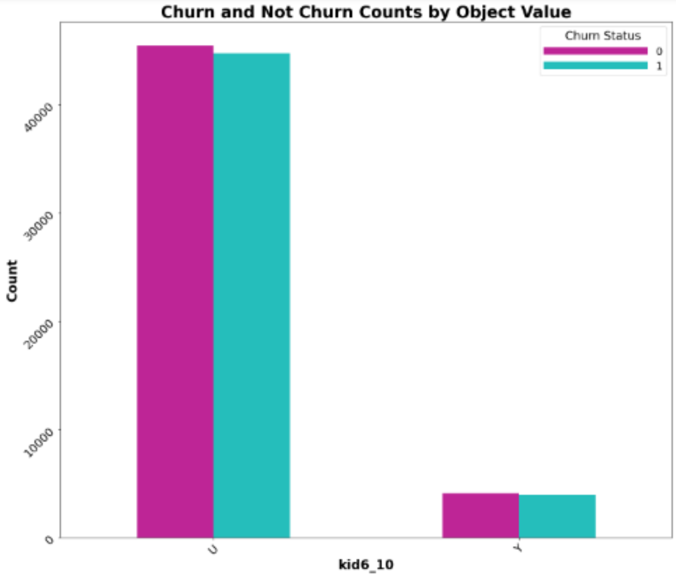

Churn and Not Churn Counts by Object Value
Churn Status
0
1
Count
0
1000
2000
3000
4000
5000
Anuradhapura
Badulla
Colombo
Galle
Gampaha
Hambantota
Jaffna
Kalutara
Kandy
Kilinochchi
Kurunegala
Matale
Matara
Millativu
Monaragala
NuwaraEliya
Polonnaruwa
Puttalam
Vaunia
area

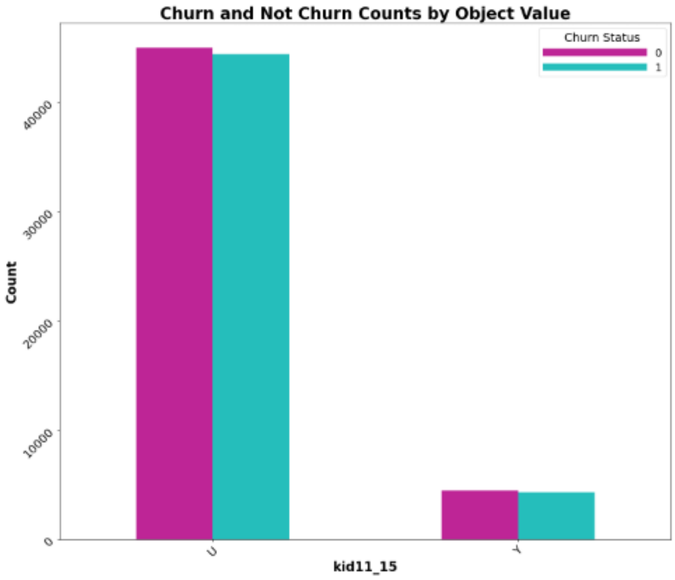

Churn and Not Churn Counts by Object Value
Churn Status
0
1
Count
0
5000
10000
15000
20000
25000
30000
35000
N
T
U
Y
dualband

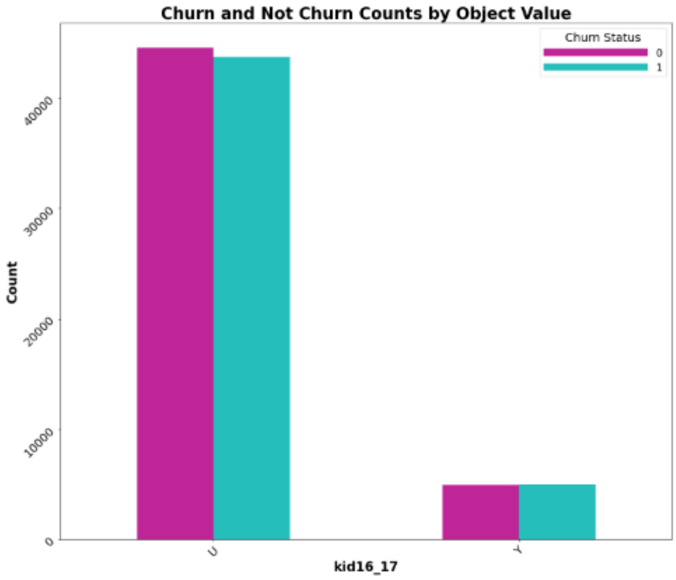

Churn and Not Churn Counts by Object Value
Churn Status
0
1
0
10000
20000
30000
40000
N
R
refurb_new


Churn and Not Churn Counts by Object Value
Churn Status
0
1
Count
0
5000
10000
15000
20000
25000
30000
35000
40000
WC


Churn and Not Churn Counts by Object Value
Churn Status
0
1
Count
0
10000
20000
30000
40000
U
Y
kid3_5


Churn and Not Churn Counts by Object Value
Churn Status
0
1
Count
0
10000
20000
30000
40000
U
Y
kid6_10


Churn and Not Churn Counts by Object Value
Churn Status
0
1
Count
0
10000
20000
30000
40000
U
Y
kid11_15


Churn and Not Churn Counts by Object Value
Churn Status
0
1
Count
0
10000
20000
30000
40000
U
Y
kid16_17

The second observation is analyzing churn distribution. According to the below. Chart 1, churn status is equally distributed.

```
churn_status
0    50438
1    49562
Name: count, dtype: int64
```

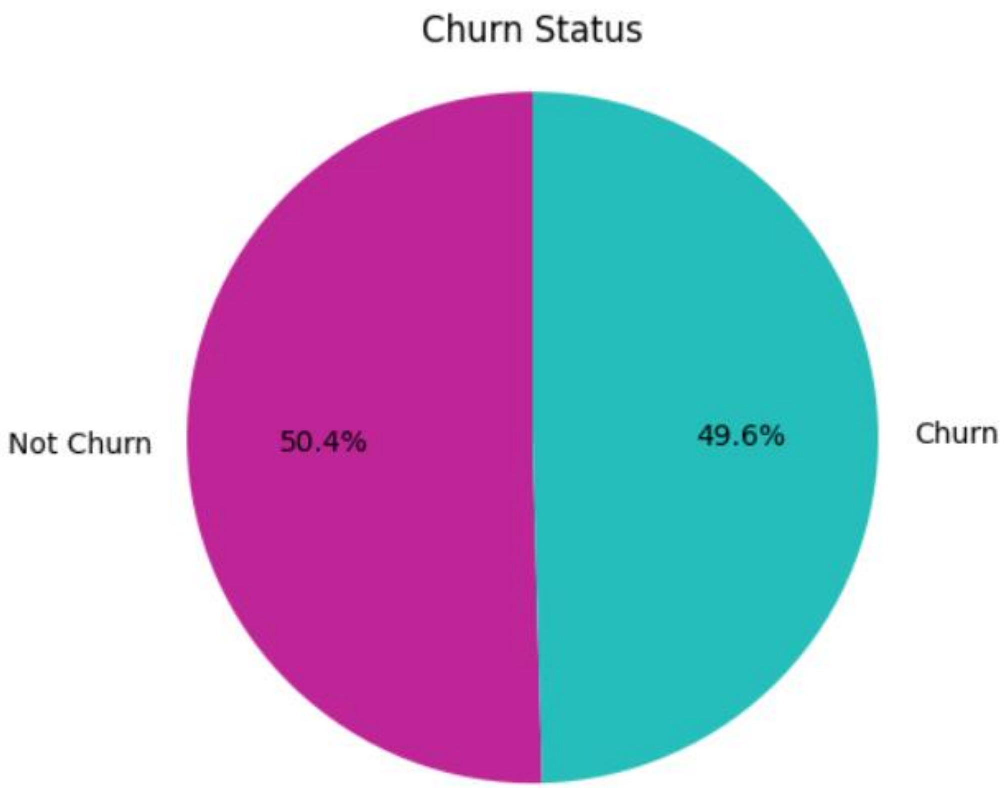


Chart 1

My subsequent analysis focusses on determining whether are there any null or missing values.[6] Handling missing values or null values is important and recommended to effective and accurate result.[7]

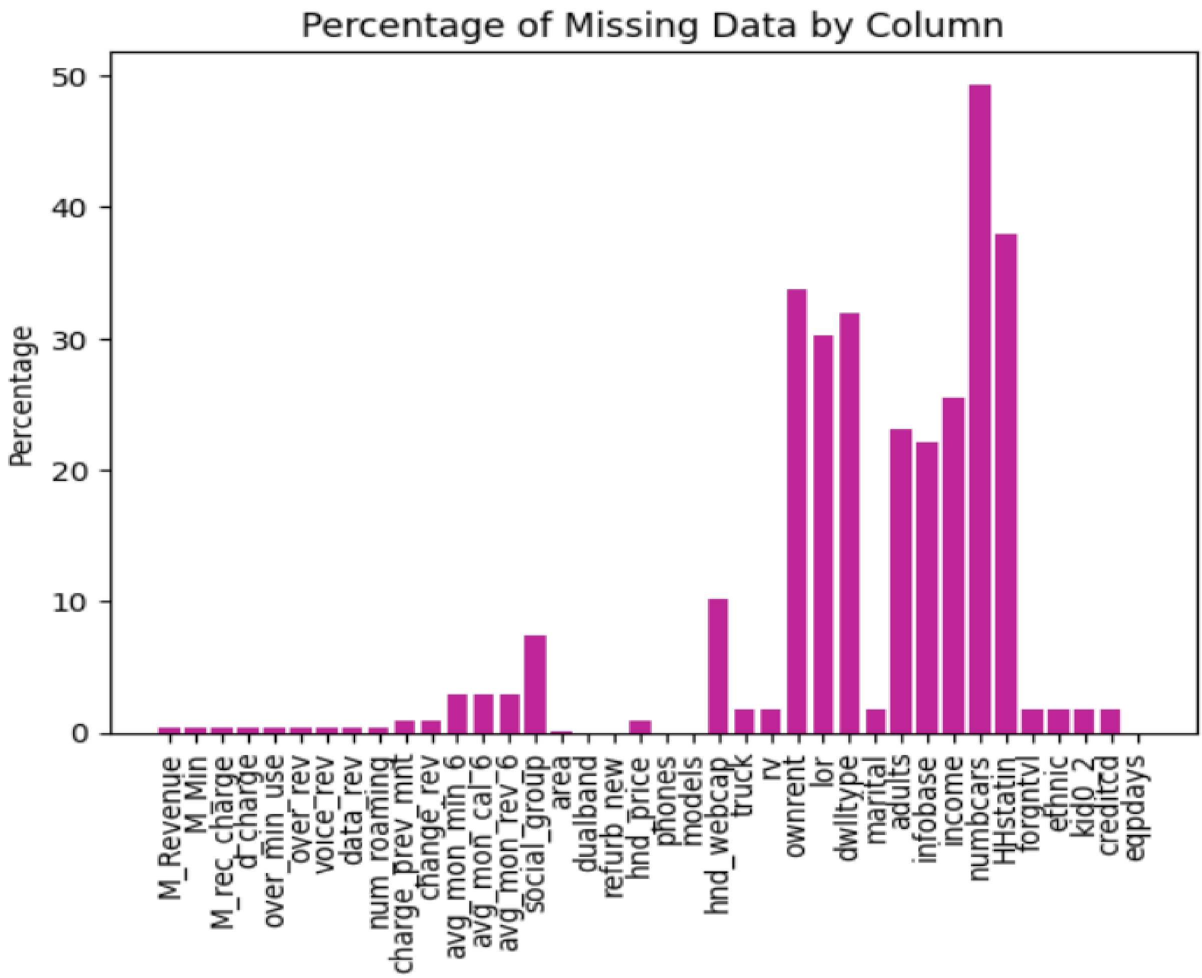


Chart 2

Chart 2 illustrates the null and missing value distribution as a percentage. According to the observation a considerable number of columns contain nearly 30 % null or missing values. I decided to fill numerical features attributes with **median** and categorical features with **mode** respectively. [8].

Next, I analyze the co-relation between numerical feature attributes and the target. According to the below graph (chart 3) all the features above the blue dotted line have (P>0.05) which conclude that there is no significant evidence of the co-relation between specific feature attributed with the target.

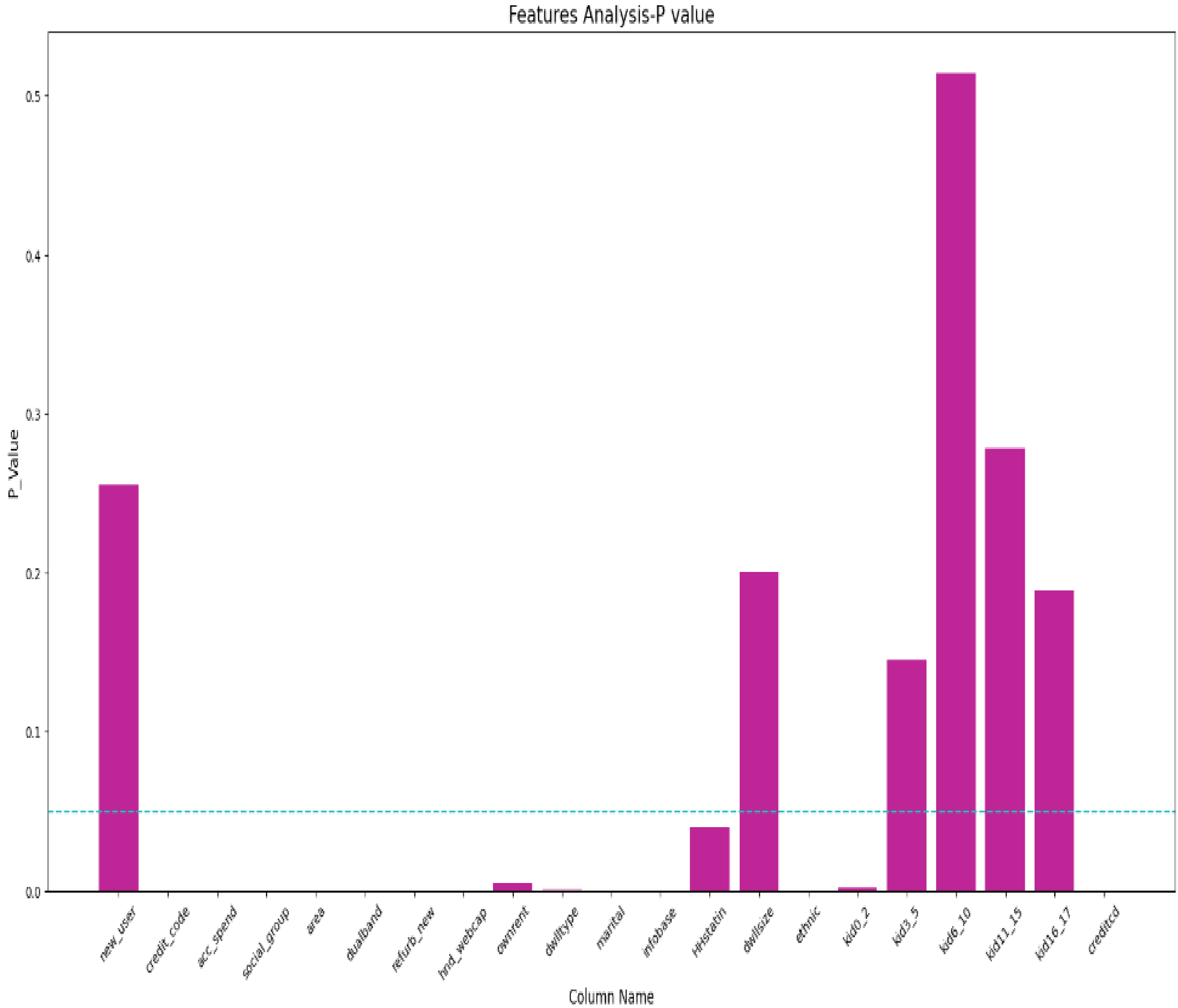


Chart 3

Same co-relation analysis performed on performed on categorical feature attributes. Charts 4.

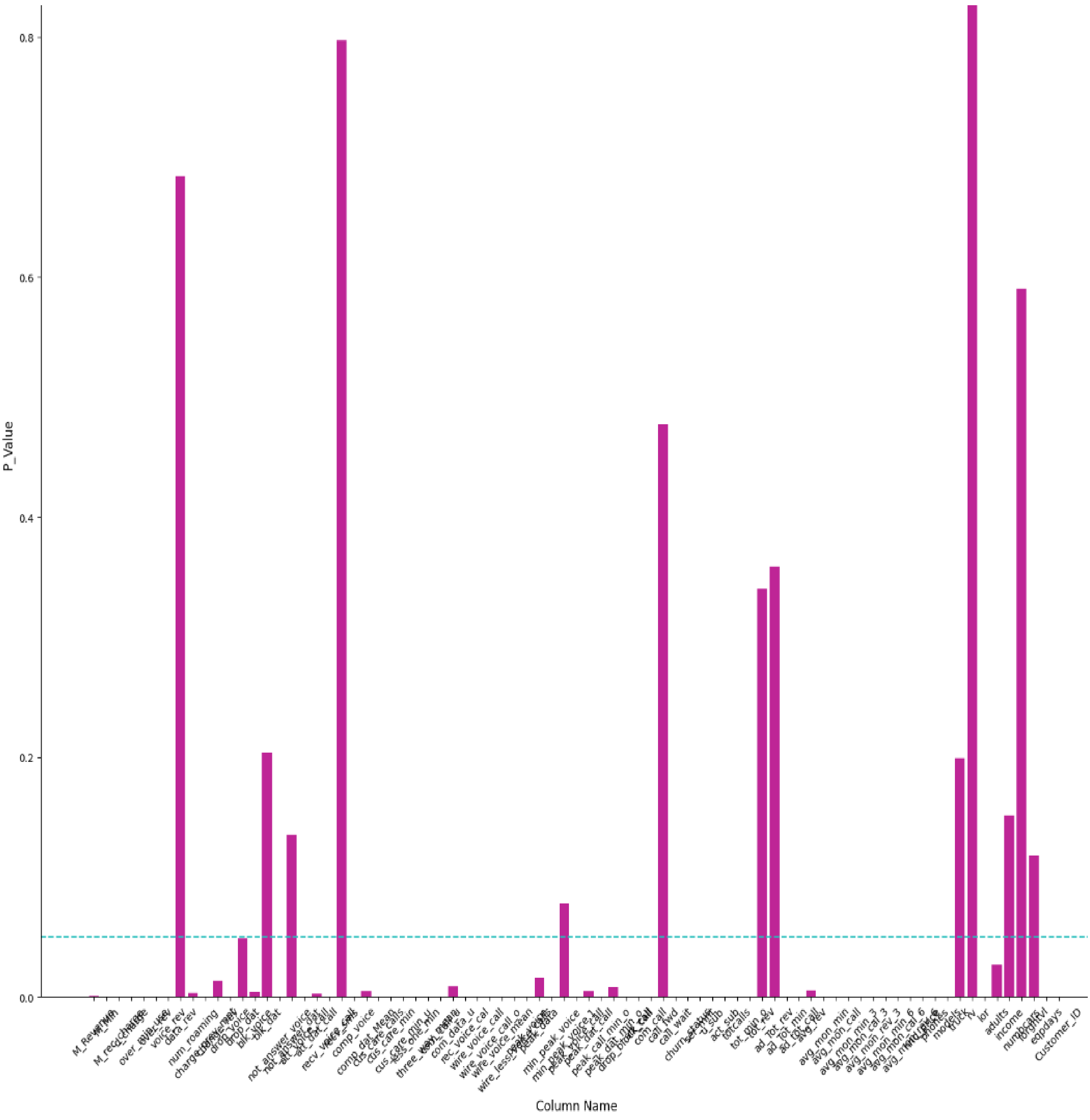


Chart 4

Outlier identification and handling is the next important factor.[9] Outlier plot (Chart 5) indicates Skewed distribution in some columns. Data transformation required. Some columns, as example “totcalls”, “tot_min_o”, “ad_Tot_min”, “avg_rev” etc. has higher values of outliers which significantly effected for the effective analysis.

I decided to handle outliers from IQR method.[10] Because IQR more focused heavily on skewed distributions, use of median.[11]

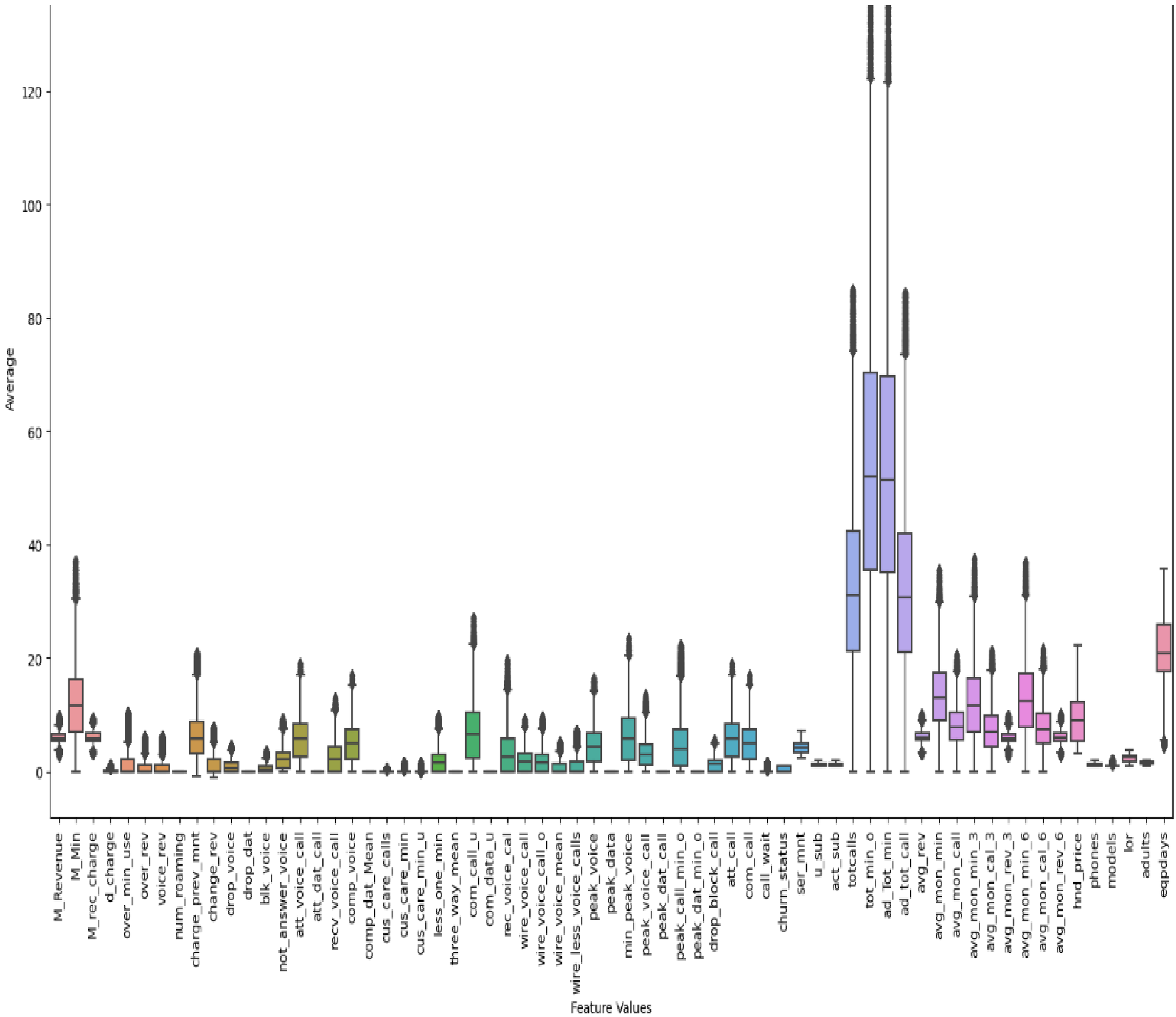


Chart 5

In feature engineering I focused cardinality identification of the categorical features. In this time, I checked the categorical features that have more cardinality variables may increase model complexity, model overfitting due to the adding more noise to the training data and computational efficiency reduction. Especially in this dataset, the number of features is relatively high and effective feature engineering is required. To effectively do feature engineering I group nonfrequent occurrences together to reduce dimensionality of the dataset.

"**Credit Code**", "**Area**" and "**Ethic**" has highest number of cardinalities, I performed analyzing of above categorical features separately and group each label by threshold of 5% (0.05). Then I replaces categories with occurrences below the threshold as "non-frequent" as below in chart6,7, and 8.

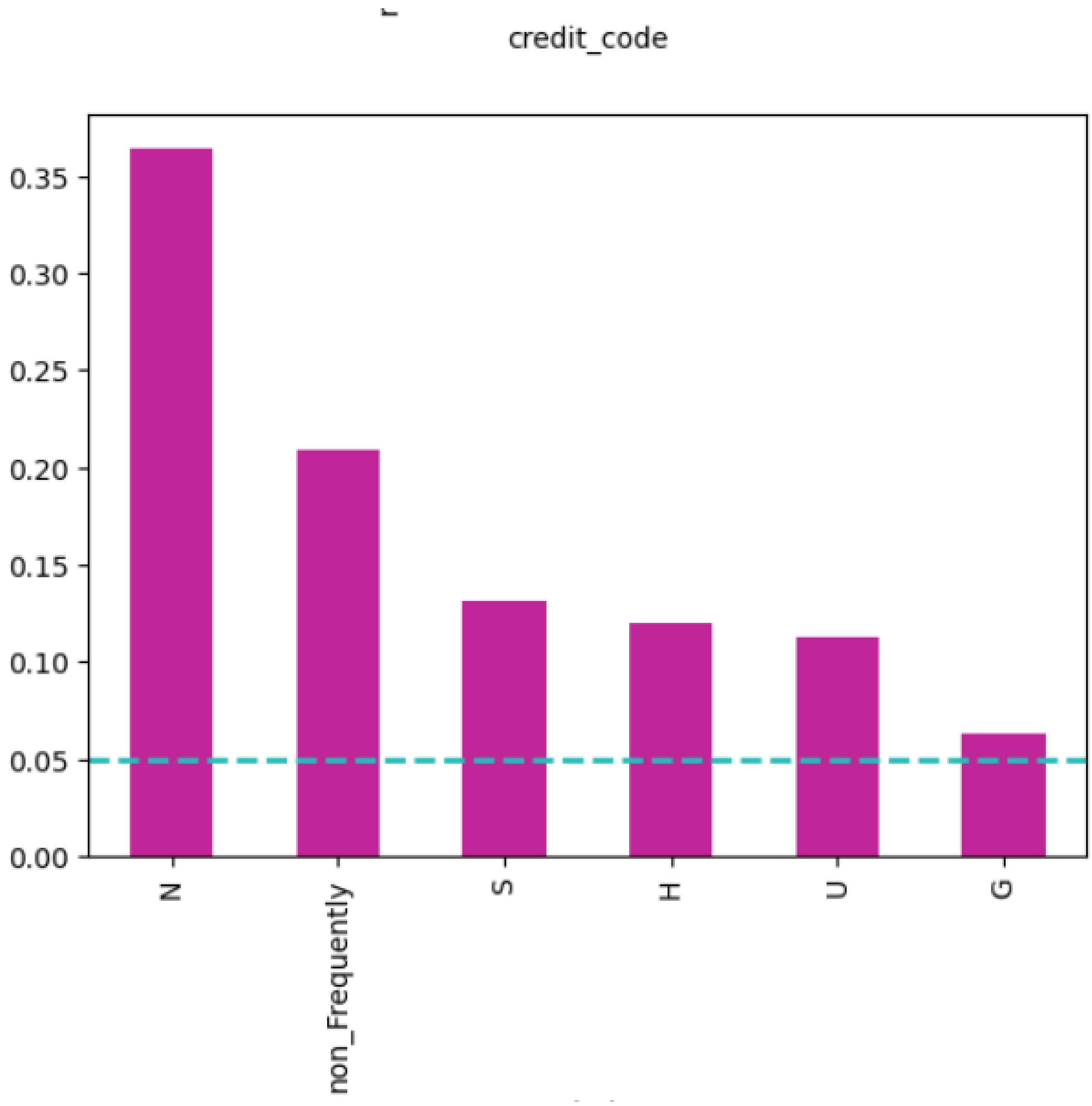


Chart 6

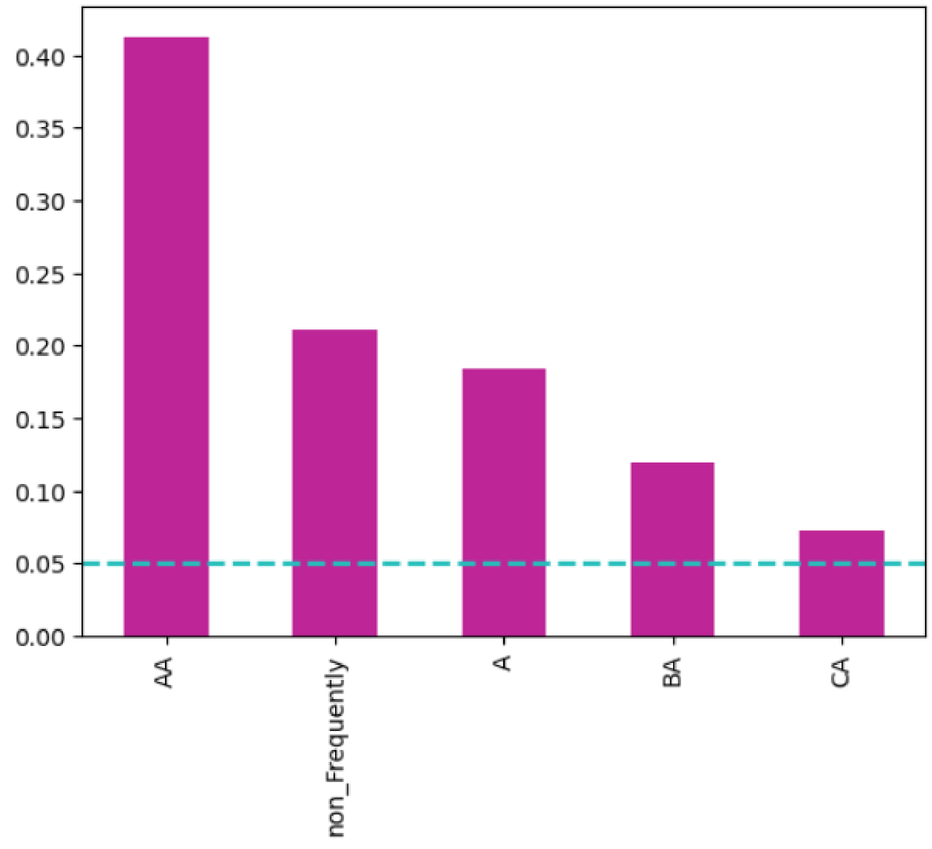


Chart 7

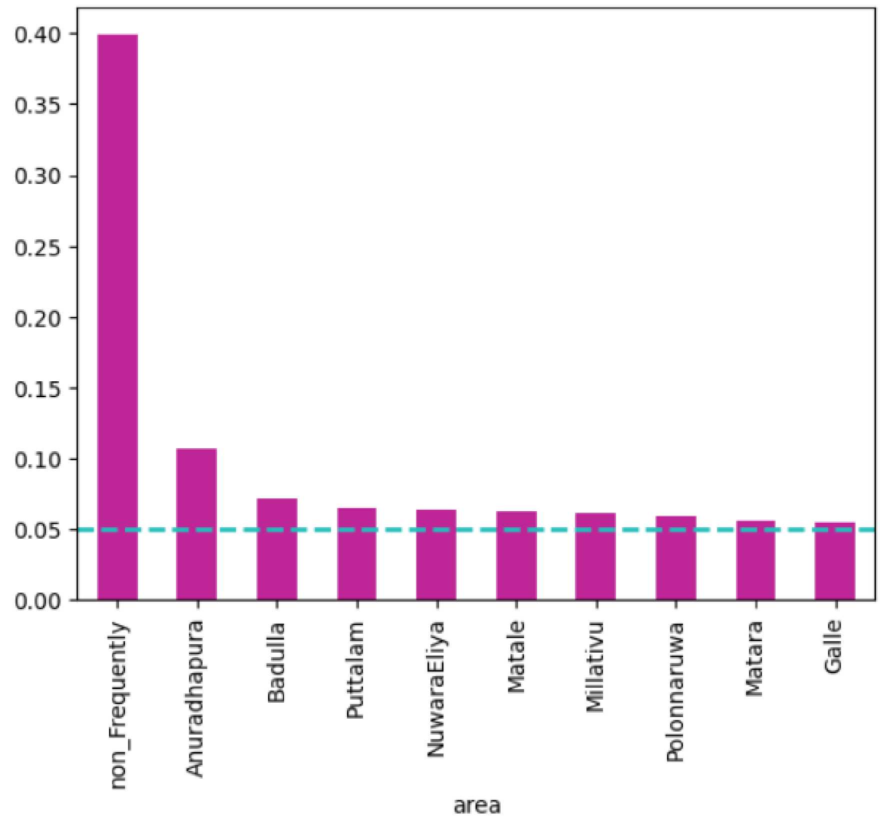


Chart 8

Next step is encoding categorical variables, In the above dataset categorical variables diversified in many different aspects. I decided to use count encoding, order integer encoding and mean encoding comprehensively to handle the above scenario.

This combination of encoding represents the frequency, ordinal relationship and the effect to the target variable which gives opportunity to enhancement of better training and prediction. [12] (Chart 9)

charge_prev_mnt	change_rev	...	refurb_new_meanEncode	hnd_webcap_meanEncode	ownrent_meanEncode	dwlltype_meanEncode	marital_meanEncod
10.665365	3.916312	...	0.500653	0.495556	0.507979	0.505171	0.53219
6.403124	2.985381	...	0.558244	0.495556	0.507979	0.505171	0.53219
7.549834	3.536948	...	0.558244	0.565684	0.507979	0.505171	0.48675
5.744563	-0.487500	...	0.500653	0.495556	0.507979	0.505171	0.53219
5.024938	2.034699	...	0.500653	0.495556	0.507979	0.505171	0.50133
...	...	...	...	...	...	...	..
1.224745	0.000000	...	0.558244	0.495556	0.507979	0.505171	0.48675
8.660254	0.000000	...	0.500653	0.495556	0.507979	0.519530	0.53219
15.288885	4.904590	...	0.500653	0.495556	0.507979	0.505171	0.53219
5.700877	-0.270000	...	0.500653	0.495556	0.507979	0.505171	0.50133
2.958040	2.021138	...	0.500653	0.495556	0.507979	0.505171	0.53219

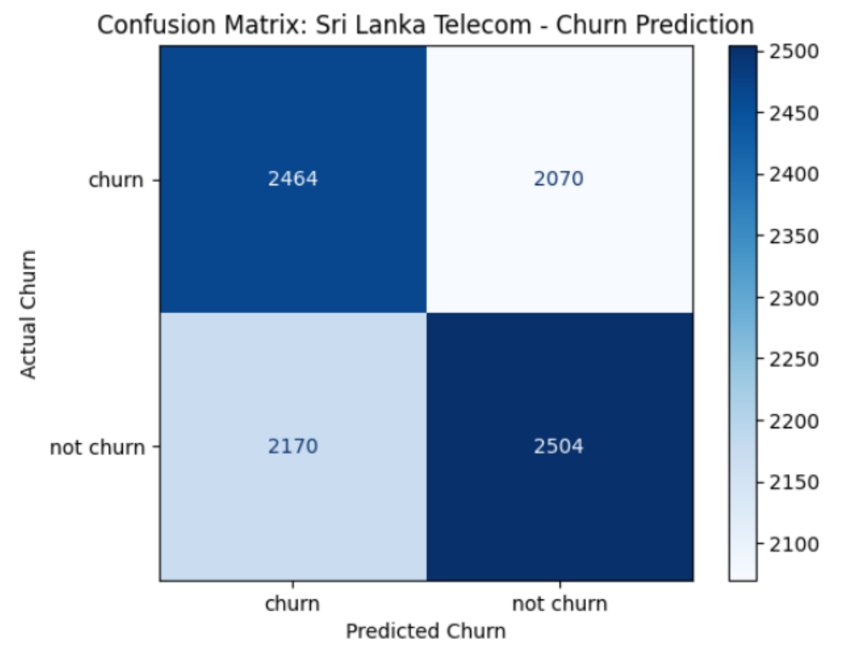
Chart 9

Classification

After successfully completing encoding all features were scaled using Standard Scaler to ensure that equal contribution, standardize high variant features like ‘change’, ‘min_usage’ etc. This ensures the feature contribution equality to the result.

Table 1

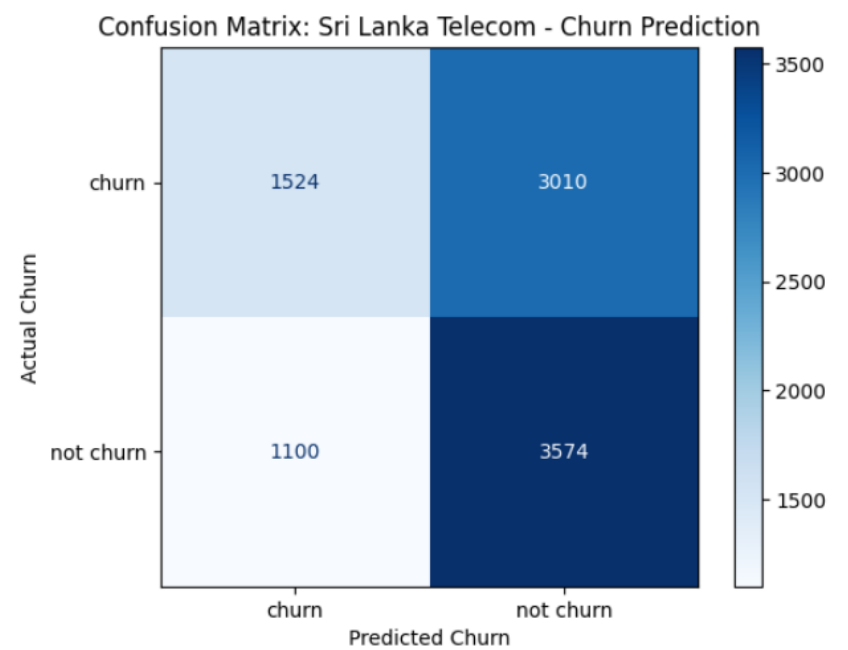
Classifier	KNN classifier
Parameters	metric="manhattan",n_neighbors=5,weights="distance"
Test Accuracy	0.54
	Accuracy Score : 0.5395308427454387 Percision Score : 0.5474420638390906 Recall Score : 0.535729567821994 F1 Score : 0.541522491349481

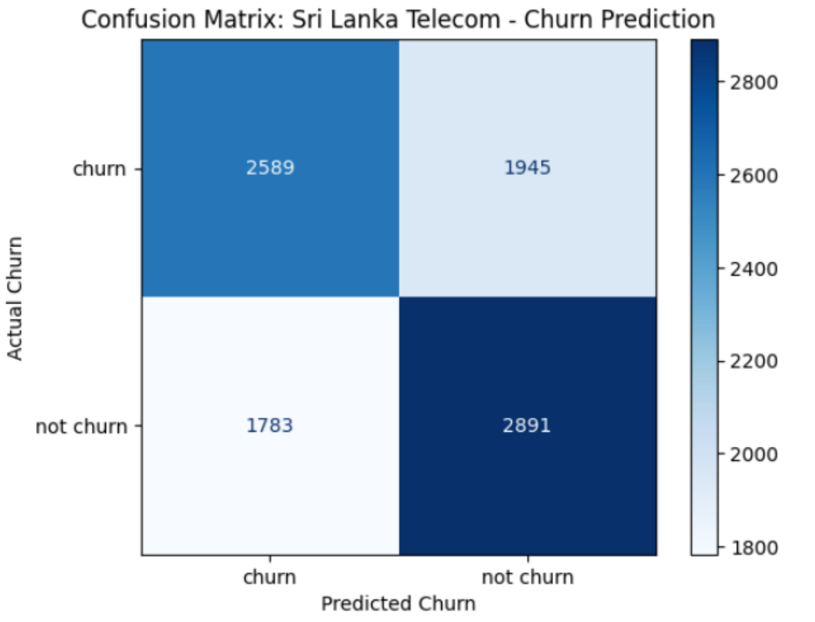
2.

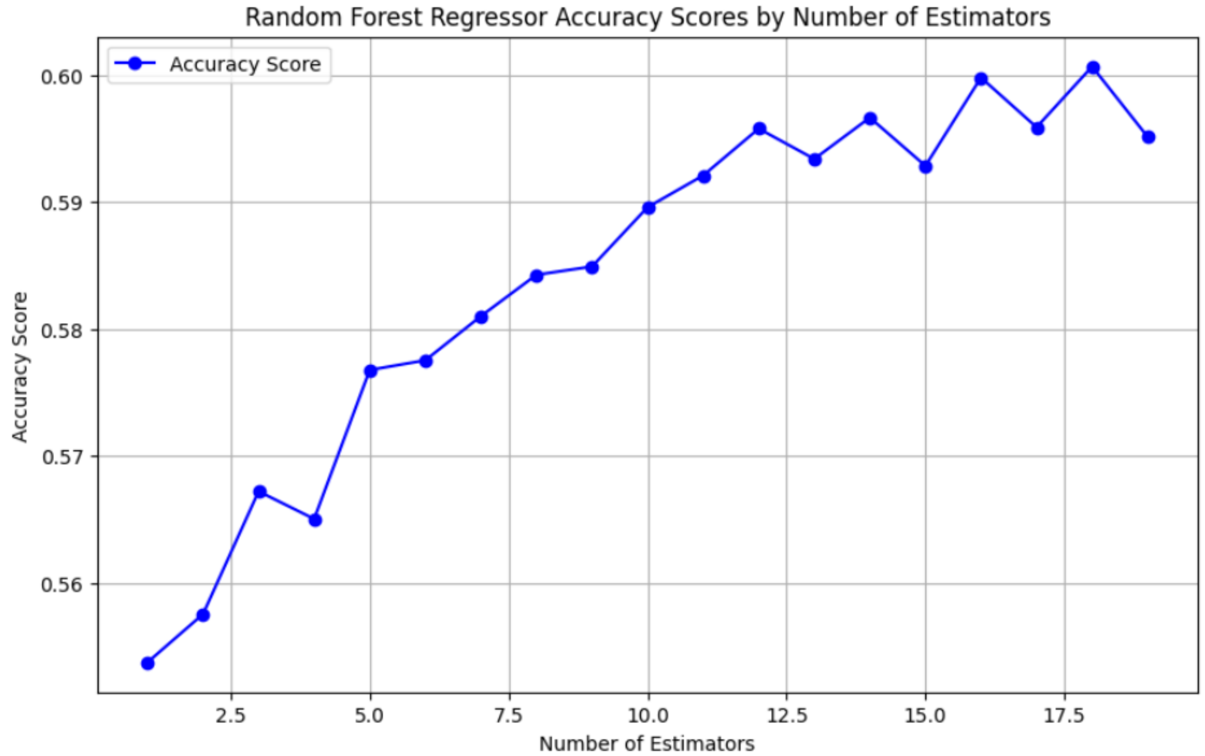
Classifier	Naive bais
Parameters	'var_smoothing': np.logspace(0, -9, num=100) cv=5, n_jobs=-1, verbose=2, scoring='accuracy
Test Accuracy	0.55

	`Accuracy Score : 0.5536490008688097`
	`Percision Score : 0.542831105710814`
	`Recall Score : 0.764655541292255`
	`F1 Score : 0.6349262746491383`

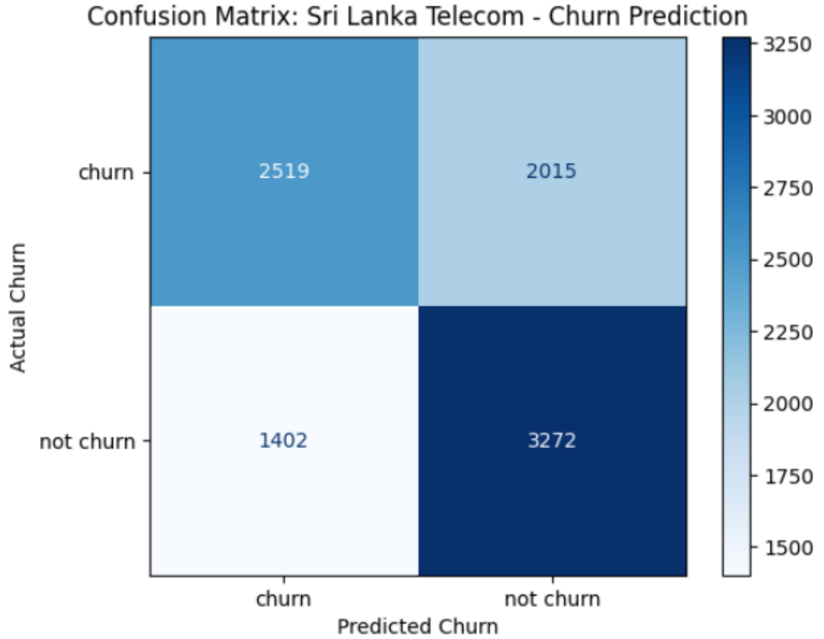
	

3.

Classifier	Random Forest Classifier
Parameters	n_estimators=18
Accuracy	0.60
	Accuracy Score : 0.6051346655082537 Percision Score : 0.597808105872622 Recall Score : 0.618528027385537 F1 Score : 0.6079915878023134
	

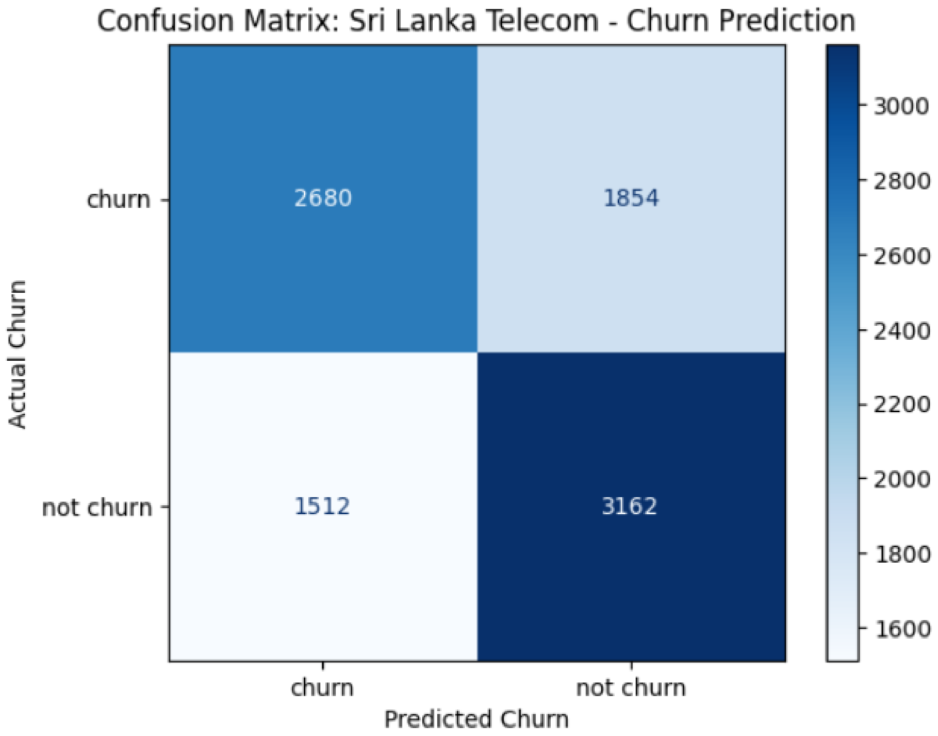


4.

Classifier	Gradian Bhoost
Parameters	n_estimators=100, learning_rate=0.1, max_depth=3, random_state=42
Accuracy	0.63
	Accuracy Score : 0.6289096437880104 Percision Score : 0.6188764895025535 Recall Score : 0.7000427899015832 F1 Score : 0.6569621523943379
	

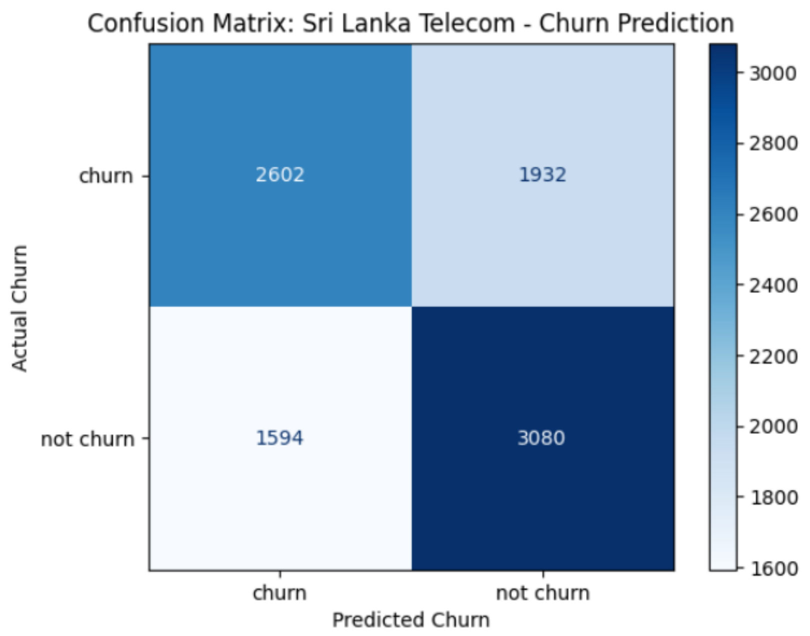
5.

Classifier	Light Gradiant Boost
Parameters	'boosting_type': 'gbdt', 'objective': 'binary', 'metric': 'binary_logloss', 'num_leaves': 31, 'learning_rate': 0.05, 'feature_fraction': 0.9, num_boost_round=100
Test Accuracy	0.63
	Accuracy Score : 0.6344483058210252 Percision Score : 0.6303827751196173 Recall Score : 0.6765083440308087

	F1 Score : 0.6526315789473685
	

6.

Classifier	Ada Boost
Parameters	DecisionTreeClassifier(max_depth=2), n_estimators=500, learning_rate=1, random_state=42
Accuracy	0.62
	Accuracy Score : 0.6170721112076455 Percision Score : 0.6145251396648045 Recall Score : 0.6589644843816 86 F1 Score : 0.6359694404294859



Classification In depth

Model Selection and Hyper parameter tuning.

KNN

KNN was my first algorithm selection which gives less accuracy compared to others in this scenario. (**Table 1).** Initial training complete using *metric="manhattan",n_neighbors=5,weights="distance"* and got accuracy around 0.54. In KNN fine tuning phase I used Grid Search [13][14]. Train and Test Score plot. (Chart 10). According to Chart best value for K Is 15. Furthermore I performed Grid Search CV which required high computational power to identify best K value. Which gives 29 as best K (Screen capture 1). KNN's accuracy score after finetuning was nearly 56%. (Screen capture 2).

```
]:  kf=KFold(n_splits=5,shuffle=True,random_state=42)
    parameter={'n_neighbors': np.arange(2, 30, 1)}
    knn=KNeighborsClassifier()
    knn_cv=GridSearchCV(knn, param_grid=parameter, cv=kf, verbose=1)
    knn_cv.fit(X_train, y_train)
    print(knn_cv.best_params_)

    Fitting 5 folds for each of 28 candidates, totalling 140 fits
    {'n_neighbors': 29}
```

Screen capture 1.

```
knn=KNeighborsClassifier(n_neighbors=29)
knn.fit(X_train, y_train)
y_pred=knn.predict(X_test)
accuracy_score=accuracy_score(y_test, y_pred)*100
print("Accuracy for testing dataset after tuning : {:.2f}%".format(accuracy_score))
```

```
Accuracy for testing dataset after tuning : 55.98%
```

Screen Capture 2.

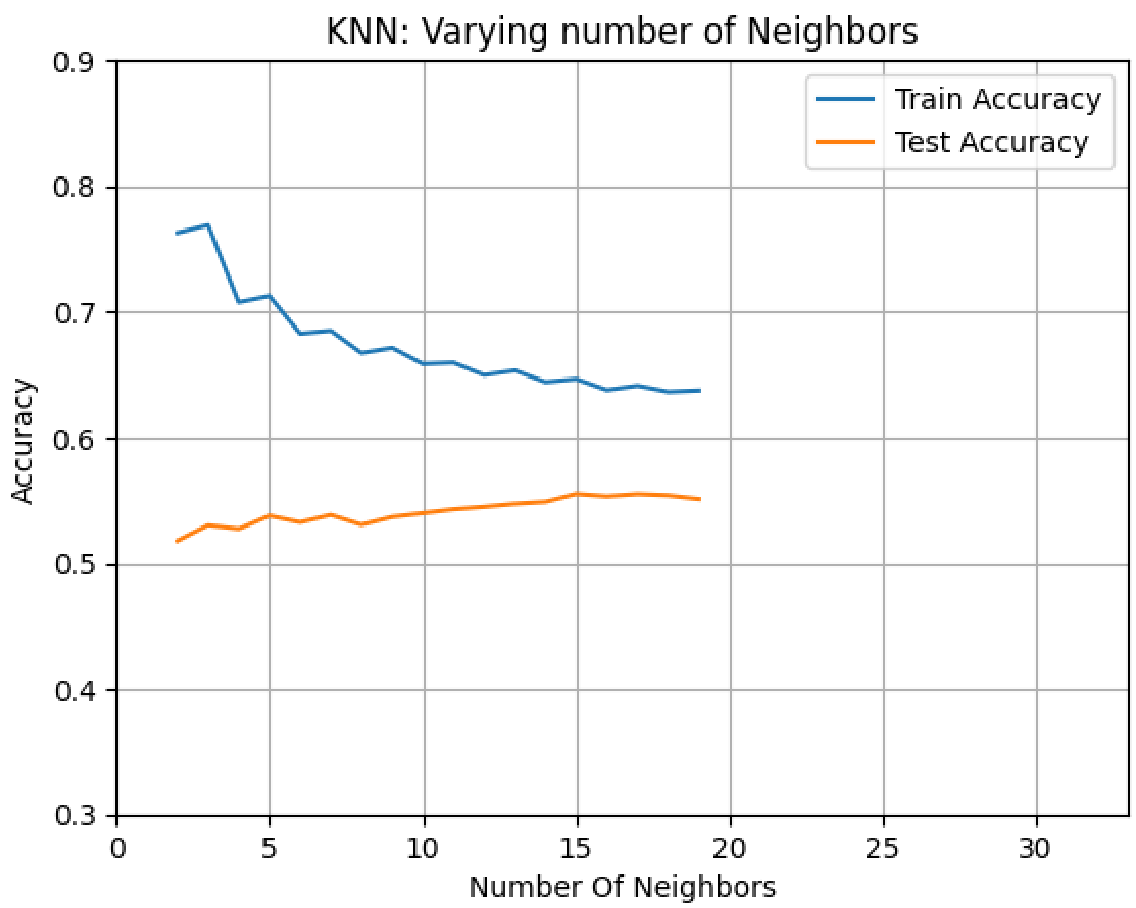


Chart 10

Light Gradient Boost

LGB is an effective, efficient, and highly scalable ML model that optimizes use of computation power and memory usage. I used the parameter setting to evaluate the LGBM model.

LGBM model has great efficiency on large data set typically with considerable categorical variables with large number of cardinalities.[15]

'num_leaves': Integer(31, 200) - *increase leaf nodes may increase over fitting but more reliable on complex scenarios* 'learning_rate': Real(0.01, 0.2, 'log-uniform'), - *leaning_rate reduce the overfitting and improve optimum model efficiency.*

'feature_fraction': Real(0.6, 1.0),

'bagging_fraction': Real(0.6, 1.0),

'max_depth': Integer(-1, 50),

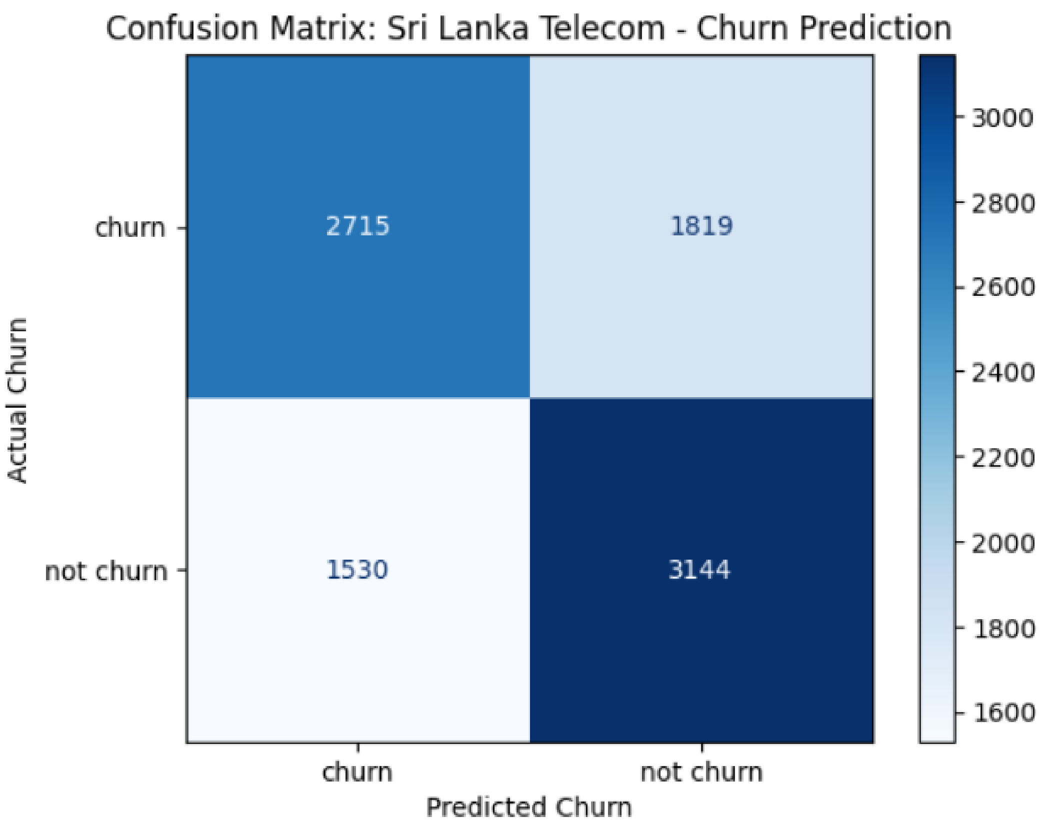
'min_data_in_leaf': Integer(20, 200),

'lambda_l1': Real(0.0, 1.0, 'uniform'),

'lambda_l2': Real(0.0, 1.0, 'uniform')

Optimum parameters -LGBM model

Best Parameters: OrderedDict([('bagging_fraction', 0.6056173284435363), ('feature_fraction', 1.0), ('lambda_l1', 0.45066439949434706), ('lambda_l2', 0.21545919864891155), ('learning_rate', 0.098331452911655), ('max_depth', 50), ('min_data_in_leaf', 200), ('num_leaves', 31)])




```
Accuracy Score     : 0.6362945264986968
```

```
Precision Score     : 0.6334878097924642
Recall Score        : 0.6726572528883183
F1 Score            : 0.652485213240635
```

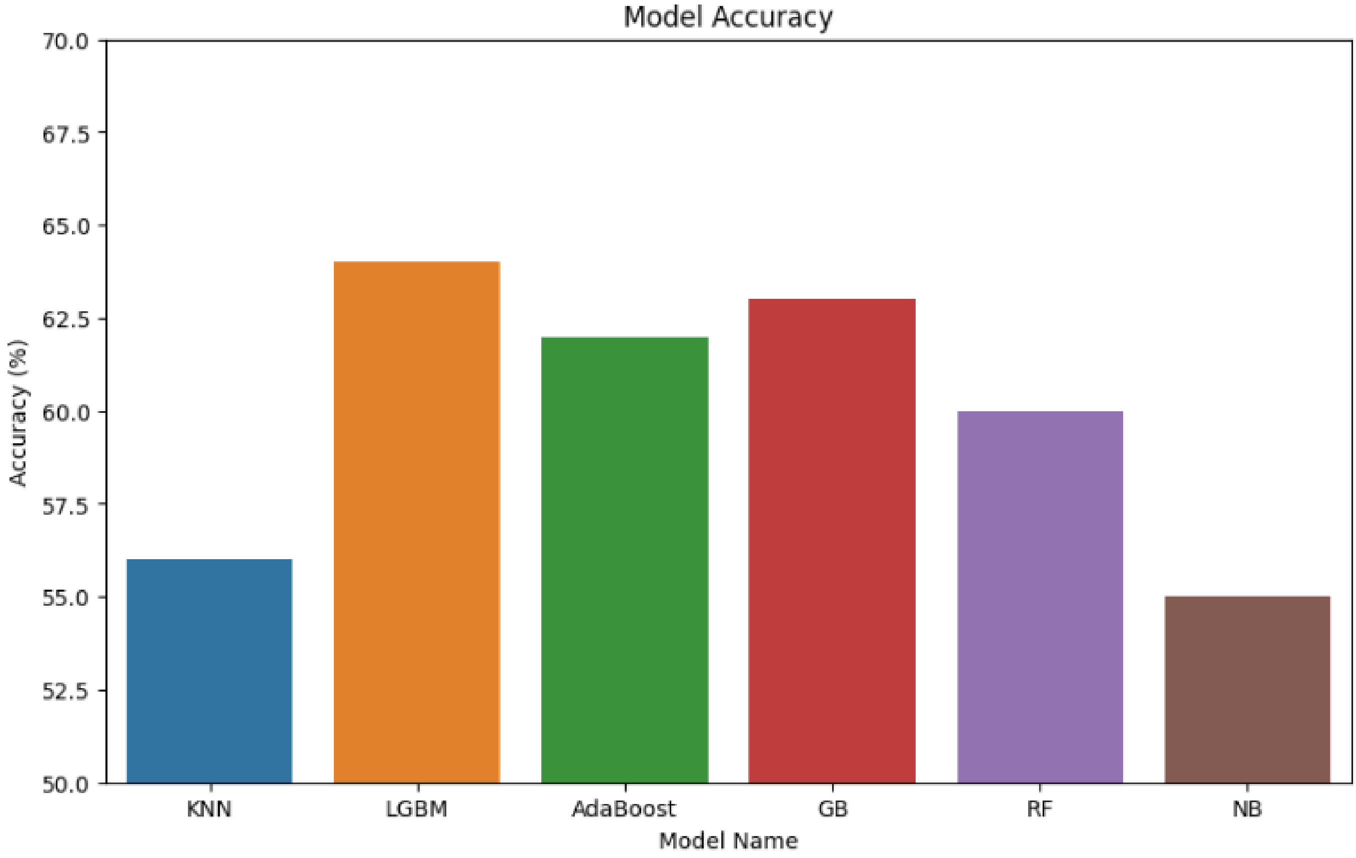


Chart 12

Feature importance for LGBM model with higher accuracy of 64 %

Features highly to determine the churn status (Chart 11).

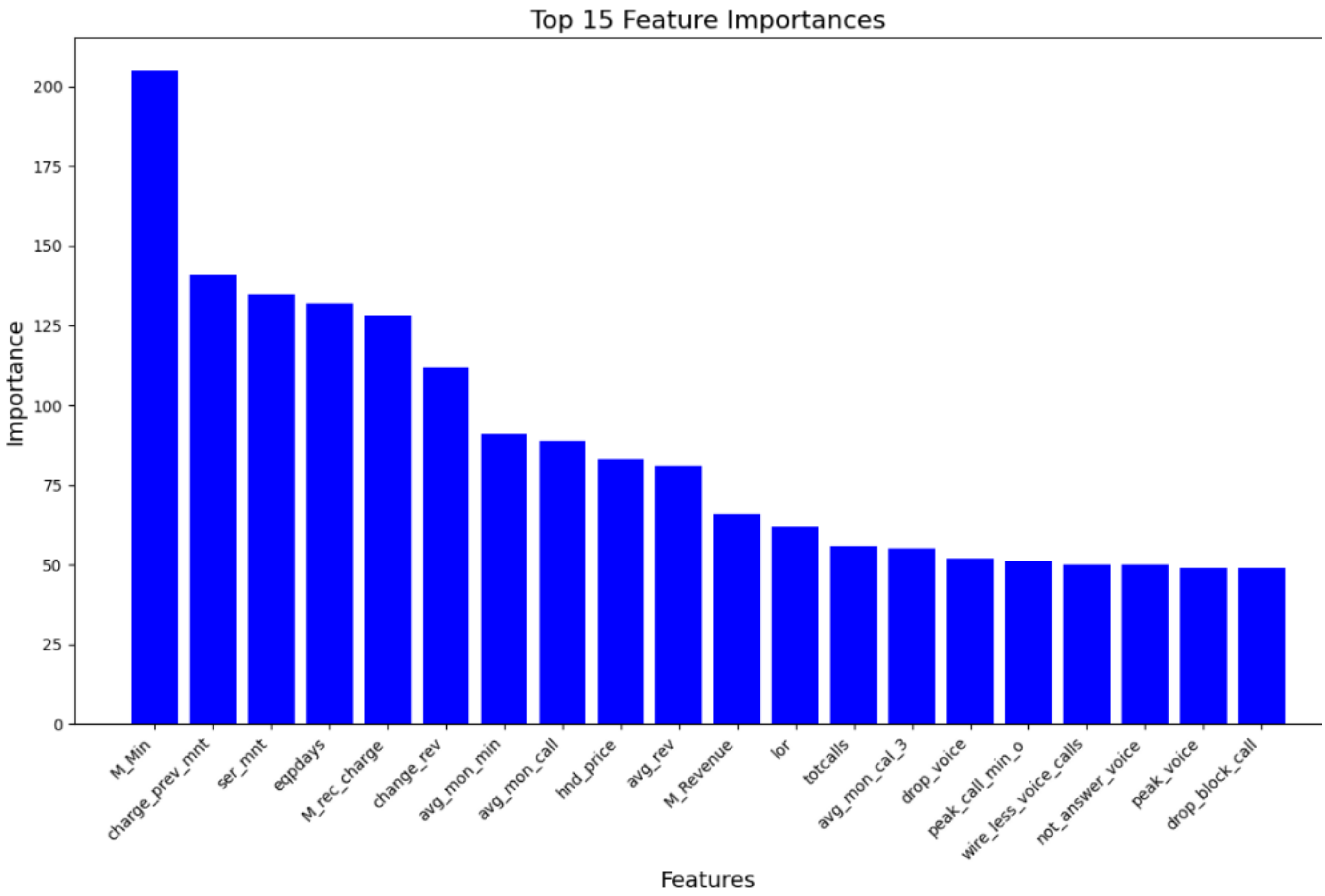


Chart 11

Conclusion

- Features Identification

After analyzing the above important features attributes to the target churn variable, below features need to be more considerable.

Calling Duration, Charge, Service Uptime, Current CPE age, Monthly recurring charge, Monthly revenue, Average month min, current_equipment_price, Resistance distance to the distribution point, Call and Data Drops, No. if bound outbound calls, wireless to wireless voice calls.

Sri Lanka Telecom management need attention on above feature attributes under various identified divisions, here I listed the ways these features can categorized.

<u>Technical Parameters- [Technical Department]</u>

Drop Calls -High frequency of drop called may issue of the Network issue can cause customer dissatisfaction.

Drop Block calls – This feature indicates the congestion of the network may cause customer dissatisfaction.

Service Uptime - Lower service Uptime is in the sense that higher outages, that cause dissatisfied customers.

CPE age- Old CPEs (customer premises equipped) may cause service failures and limited opportunities to upgrade services.

Wireless Voice calls – Voice over wireless may significantly affect to increase customer satisfaction.

Sales and Marketing

Minutes of Usage – Higher minutes indicate that customers use the service more. A higher level of attention, specific promotions and customer monitoring are needed.

Charge per minute; This discusses pricing strategy. Higher prices may cause higher the customer churn.

Customer Revenue; More specific packages and customize service options base on customer revenue may increase customer protection.

- Model Section

I suggest the Light Gradient Boost model with Max Depth=50 and num leaves=31.

Accuracy of 64%

References

1. Vladislav L., Marius C. (2019) Churn Prediction: (2019.August.18)
2. Sulaimon O. (2015) Predicting Customer Churn and Retention Rates in Nigeria's Mobile Telecommunication Industry Using Markov Chain Modelling (2015)
3. Haoyi Xiong MPaaS: Mobility prediction as a service in telecom cloud https://link.springer.com/article/10.1007/s10796-013-9476-z
4. Amel Salem Omer, and Dereje Hailemariam Woldegebreal, (2022) Review of Markov Chain and Its Applications in Telecommunication Systems. [online]. 1 [Accessed 26 May 2022].
5. Abdelrahim Kasem Ahmad, , Assef Jafar, and Kadan Aljoumaa, (2019) Customer Churn Prediction in Telecom Using Machine Learning in Big Data Platform. Journal of Big Data [online]. 1 [Accessed 14 May 2024].
6. Swj Nijman, , Am Leeuwenberg, and I Beekers, (2019) Missing Data Is Poorly Handled and Reported in Prediction Model Studies Using Machine Learning: A Literature Review. Journal of Clinical Epidemiology [online]. 1 [Accessed 14 May 2024].
7. Zulfikar Irham,, , and , (2021) Telco Customer Churn Prediction Using Machine Learning and Deep Learning. Media [online]. 1 [Accessed 17 May 2024].

8. Satyam Kumar,, , and , (2022) 7 Ways to Handle Missing Values in Machine Learning. Media [online]. 1 [Accessed Jul 24, 2020].
9. Abir Smiti, , , and , (2023) A Critical Overview of Outlier Detection Methods. Science Direct [online]. 1 [Accessed November 2020]
10. Amerah Alabrah, (2023) An Improved Ccf Detector to Handle the Problem of Class Imbalance with Outlier Normalization Using Iqr Method. Mdpi [online]. [Accessed 2024 May]
11. H. P. Vinutha, and B. M. Sagar, (2018) Detection of Outliers Using Interquartile Range Technique From Intrusion Dataset. Springerlink [online]. [Accessed 2024 May].
12. Ivan Lopez-arevalo, and Edwin Aldana-bobadilla, (2020) Settingsorder Article Reprints Open Accessarticle a Memory-efficient Encoding Method For Processing Mixed-type Data on Machine Learning. Mdpi [online]. [Accessed 2024 May].
13. Pirjatullah, and Dwi Kartini, (2021) Hyperparameter Tuning Using Gridsearchcv on the Comparison of the Activation Function of the Elm Method to the Classification of Pneumonia in Toddlers. Ieee Xplore Logo [online]. [Accessed 2024 May].
14. Ravi Singh. (2023) Choosing the Best: A Comparison between GridSearchCV and RandomizedSearchCV Available from: https://www.linkedin.com/pulse/choosing-best-comparison-between-gridsearchcv-ravi-singh/ [Accessed 16 May 2024].
15. Afek Ilay Adler, (2022) Feature Importance in Gradient Boosting Trees with Cross-validation Feature Selection. Mdpi [online]. [Accessed May 2024].