CUSTOMER CHURN ANALYSIS

Masterclass in Business Analytics-Group S14

DEV UPADHYAY
200644879

Contents

Abstract :

Customer churn is a big concern for organizations in the modern-day due to increasing competition, the significance of marketing strategies, and customers' heightened knowledge. Attrition from one telecom service provider to another happens when competitor organisations offer clients more appealing telecom plan pricing or superior services when signing up. In the telecom industry, churn has become an integral aspect of planning and strategic decision making. Companies in the competitive and fast evolving telecom market must devise new tactics to combat churn. This project is an effort to estimate customer churn for a telecommunications firm based on a data set including numerous variables describing the qualities of the sector and other elements deemed significant when dealing with telecom consumers. This data collection has 100 variables and around 100,000 records. Using machine learning techniques such as linear regression, logistic regression, and random forest, the telecom industry can anticipate user turnover. According to the study's findings, the accuracy rate of customer churn prediction is ~0.64 percent and area under the curve is ~0.65. In this predictive study, we used logistic regression, Gaussian Naive Bayes, Random forest, XG Boost classifier, and artificial neural networks (ANN). Additionally, we utilised one-hot encoding to convert categorical data to numeric ones and the log transformation approach to manage outliers. Following is our proposed solution flowchart :

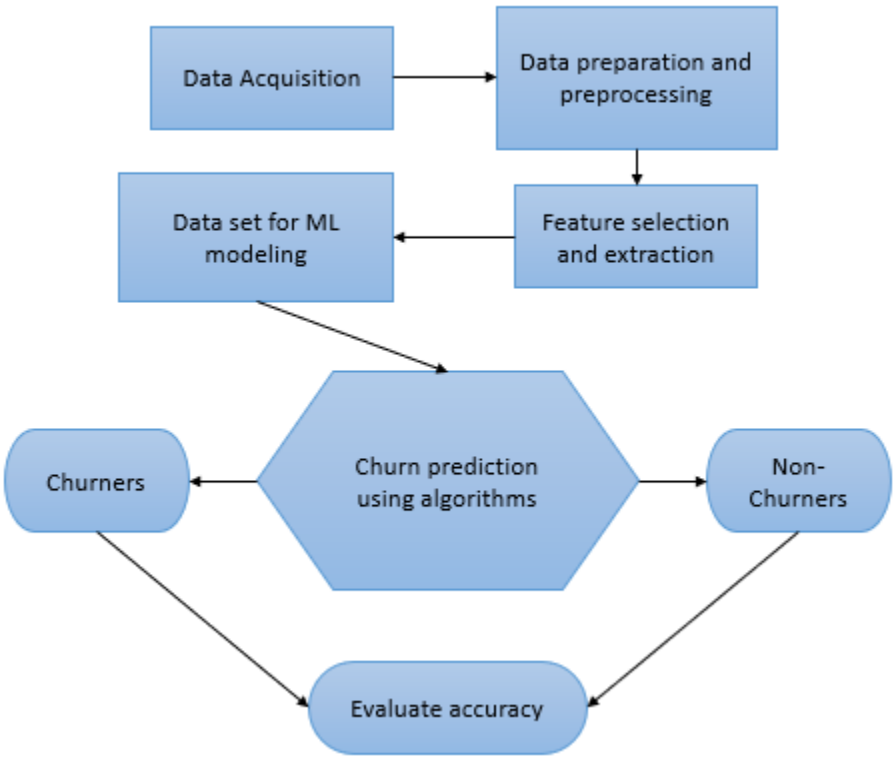

Fig 1: Flowchart of proposed solution

Introduction :

Telecommunications has become a major industry in industrialised nations. The rise of technology and operators increased competition. As a result, businesses understand the value of maintaining existing clients. Service providers must now estimate customer attrition to minimise customer churn. Many studies have shown that machine learning is particularly effective in this case. For a decade, several studies on churn prediction have used statistical and machine learning techniques. This section covers the significant churn prediction papers in the telecom sector.

According to Alwin (2018), when service providers understand why clients leave, they may improve their offerings to meet their needs. Analyzing prospective consumers' past behavior might significantly reduce churn. The system was built using logistic regression and neural networks. Finally, a comparison analysis is performed to choose the most favorable model and examine it with precision and consistency. The research suggested using the C5.0 algorithm of the decision tree to control churn. The above model was found to be the most accurate, with an AUC of 0.888.

Nigam et al. (2019) used machine learning to predict client churn. Historically, several machine learning techniques, such as Decision trees, SVM, Logistic Regression, SVM, NN, etc., have been used to predict customer churn. Machine learning is the process of designing algorithms that can create predictions and learn from data. Deep neural networks are a potent technique for forecasting telecom customer churn. Using numerous concept hierarchies, we may construct a model that corresponds to our data, so enhancing its performance. This study used a multi-layer artificial neural network (ANN) to predict telecom customer churn. The study's recommended model has a sensitivity of 85%, hence the outcomes are positive.

According to Talwar and Dahiya, one of the most important components of contemporary telecom CRM systems is customer churn prediction (2015). This article covers a contemporary method to churning prediction using machine learning. This paper assesses prevalent ML techniques for identifying customer churn patterns in the telecom industry. The current study aims to develop alternative hybrid methodologies in which ensemble classification systems are commonly coupled with preprocessing techniques. The study demonstrates that machine learning classifiers are successful if sufficient feature engineering effort is applied by humans. Using machine learning, an automated churn prediction system will be developed.

The report is organised such that the business need, methodology employed, and analysis of the data collected are all easily understood, followed by a conclusion and any research constraints.

Business Problem :

Globally, the sector of telecommunications is gradually becoming one of the most significant. Competition has risen as a direct consequence of the growth of technical capabilities and the increase in the number of operators. For telecom businesses to survive in this very competitive market, a variety of tactics targeted at generating substantial amounts of revenue have been created. To compete in this

highly competitive environment, businesses are using a variety of techniques. Three primary revenue-generation techniques have been proposed:

(1) get new subscribers

(2) cross-sell additional features to existing customers

(3) keep customer loyalty active.

The third technique has shown to be the most lucrative, maintaining current customers costing far less than obtaining new ones, and is also deemed simpler than upselling. Customer turnover is a major issue in highly competitive markets. Churn is the number of customers that stop purchasing from a company. In the telecommunications industry, customer churn happens when a customer stops using services such as voice, SMS, data, and mobile money.

- Non-renewal
- Cancellation

Predicting clients that are likely to quit the firm may be a significant source of extra money if done early. To increase the length of time that customers stay loyal to a firm, companies need to lower the possibility of consumer defection. Machine learning algorithm approaches assist telecom companies in defending against churn. Silent churn is difficult to predict since customers may leave in the near future. Because established customers are more valuable than new ones, decision-makers and advertisers should strive to decrease turnover. The following benefits are a direct outcome of retaining more customers:

1. There are savings involved with acquiring new consumers.
2. Value of the Net Promoter Score is Enhanced.
3. increased financial gain possibilities
4. Prosperity over the extended period.

At order to handle the issue of customer turnover in a big Chinese telecom firm that services around 5.23 million subscribers, He Y, He Z, and Zhang D presented a model for prediction that was based on the Neural Network algorithm. The total accuracy rate served as the benchmark for the accuracy of the predictions, and it was 91.1 percent accurate.

Idris A, Khan A, and Lee YS came up with a method based on genetic programming and AdaBoost to simulate the problem of churn in the telecommunications industry. Two different kinds of data sets were used to test the model. One was made by Orange Telecom and the other by cell2cell. The cell2cell dataset was 89 percent accurate, while the other one was only 63 percent accurate.

Huang F, Zhu M, Yuan K, Deng EO (2015) investigated massive data turnover. The primary purpose of the research was to demonstrate that big data improved churn prediction. All of these characteristics are necessary for large data. To develop a solution, a large Chinese telecommunications business required a big data platform. The AUC evaluated Random Forest. The more information provided, the more

accurate the churn prediction model. The algorithm's weakness was that it used just a single classification technique, preventing comparisons. Given that it depends on massive data platforms to predict attrition, it is vital to large data companies.

As the customer churn rate is the major focus of this study, our primary purpose was to identify the factors that are most responsible for customer churn. This objective was accomplished since we focused on the relationship between churn and other factors. In addition, we examined the likelihood of a relationship between the independent variable and the indirect effect potential of the churn rate.

Data, EDA, and Methods:

In order to do data preprocessing and segment the data required for training and testing, we used the four-phase pipeline structure shown in Fig. 2. Although there are other tactics and ways that may be employed to finish each step, we have outlined and examined the preferred ones.

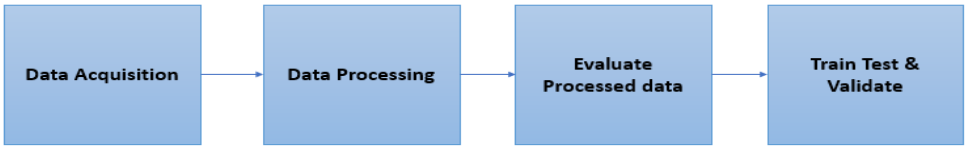

Fig 2: Pipeline of data preprocessing and data segmentation

Data used for this research can be found on the website Kaggle. In this dataset we can observe that it contains demographic, behavioral, and engagement data elements in the dataset. This dataset comprises more than 100,000 records of consumers who have used telecommunications services and includes a variety of factors that describe their characteristics. The data collection includes information on customers of telecommunications companies in 19 of the most populous cities in the United States. Each record of a customer is characterized by 100 distinct criteria which also inculudes the target variable as well. As this analysis is supervised learning so the target variable is churn, which indicates whether a customer will churn or not. Exhibit 1 allows us to see all of the variables and their respective descriptions. Below is a summary of the data frame which includes all the numeric fetures, is shown in the table below:

	rev_Mean	mou_Mean	totmrc_Mean	da_Mean	ovrmou_Mean	ovrrev_Mean	vceovr_Mean	datovr_Mean	roam_Mean	change_mou	...
count	99643.000000	99643.000000	99643.000000	99643.000000	99643.000000	99643.000000	99643.000000	99643.000000	99643.000000	99109.000000	...
mean	58.719985	513.559937	46.179136	0.888828	41.072247	13.559560	13.295062	0.261318	1.286405	-13.933818	...
std	46.291677	525.168140	23.623489	2.177619	97.296150	30.500885	30.056089	3.126531	14.711374	276.087509	...
min	-6.167500	0.000000	-26.915000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	-3875.000000	...
25%	33.260000	150.750000	30.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	-87.000000	...
50%	48.195000	355.500000	44.990000	0.247500	2.750000	1.000000	0.682500	0.000000	0.000000	-6.250000	...
75%	70.750000	703.000000	59.990000	0.990000	42.000000	14.437500	14.025000	0.000000	0.235000	63.000000	...
max	3843.262500	12206.750000	409.990000	159.390000	4320.750000	1102.400000	896.087500	423.540000	3685.200000	31219.250000	...

8 rows × 79 columns

Table 1: Data summary

Exploratory and Visual Data Analytics

When we started exploring the dataset we observed that object and float are the two primary data types for variables in the data collection, as seen in the table below:

Data Type	No of variables	Variables
object	21	'new_cell', 'crclscod', 'asl_flag', 'prizm_social_one', 'area', 'dualband', 'refurb_new', 'hnd_webcap', 'ownrent', 'dwlltype', 'marital', 'infobase', 'HHstatin', 'dwllsize', 'ethnic', 'kid0_2', 'kid3_5', 'kid6_10', 'kid11_15', 'kid16_17', 'creditcd'
float	69	'rev_Mean', 'mou_Mean', 'totmrc_Mean', 'da_Mean', 'ovrmou_Mean', 'ovrrev_Mean', 'vceovr_Mean', 'datovr_Mean', 'roam_Mean', 'change_mou', 'change_rev', 'drop_vce_Mean', 'drop_dat_Mean', 'blck_vce_Mean', 'blck_dat_Mean', 'unan_vce_Mean', 'unan_dat_Mean', 'plcd_vce_Mean', 'plcd_dat_Mean', 'recv_vce_Mean', 'recv_sms_Mean', 'comp_vce_Mean', 'comp_dat_Mean', 'custcare_Mean', 'ccrndmou_Mean', 'cc_mou_Mean', 'inonemin_Mean', 'threeway_Mean', 'mou_cvce_Mean', 'mou_cdat_Mean', 'mou_rvce_Mean', 'owylis_vce_Mean', 'mouowylisv_Mean', 'iwylis_vce_Mean', 'mouiwylisv_Mean', 'peak_vce_Mean', 'peak_dat_Mean', 'mou_peav_Mean', 'mou_pead_Mean', 'opk_vce_Mean', 'opk_dat_Mean', 'mou_opkv_Mean', 'mou_opkd_Mean', 'drop_blk_Mean', 'attempt_Mean', 'complete_Mean', 'callfwdv_Mean', 'callwait_Mean', 'totmou', 'totrev', 'adjrev', 'adjmou', 'avgrev', 'avgmou', 'avgqty', 'avg6mou', 'avg6qty', 'avg6rev', 'hnd_price', 'phones', 'models', 'truck', 'rv', 'lor', 'adults', 'income', 'numbcars', 'forgntvl', 'eqpdays'

Table 2: Data type of all the features

churn	0	1
area		
ATLANTIC SOUTH AREA	0.510274	0.489726
CALIFORNIA NORTH AREA	0.478945	0.521055
CENTRAL/SOUTH TEXAS AREA	0.521517	0.478483
CHICAGO AREA	0.504766	0.495234
DALLAS AREA	0.512532	0.487468
DC/MARYLAND/VIRGINIA AREA	0.539668	0.460332
GREAT LAKES AREA	0.523553	0.476447
HOUSTON AREA	0.524994	0.475006
LOS ANGELES AREA	0.501808	0.498192
MIDWEST AREA	0.540971	0.459029
NEW ENGLAND AREA	0.482575	0.517425
NEW YORK CITY AREA	0.499730	0.500270
NORTH FLORIDA AREA	0.480000	0.520000
NORTHWEST/ROCKY MOUNTAIN AREA	0.430915	0.569085
OHIO AREA	0.536245	0.463755
PHILADELPHIA AREA	0.493459	0.506541
SOUTH FLORIDA AREA	0.466387	0.533613
SOUTHWEST AREA	0.489898	0.510102
TENNESSEE AREA	0.528681	0.471319

In the second phase of data processing while checking the missing values in the data we observed that there 10 variables with more than 10% of data is missing, where numbcars has ~49%, dwllsize & HHatatin has ~38%, ownrent has ~33%, etc, of data is missing, as shown in exhibit 2.

After the data has been adequately processed, it must be visualised so that valuable insights may be derived from it that may be relevant for future research. From the following examination of the area and churn variable table 2 and bar plot fig 3, we can conclude that customer attrition is comparable across all regions and non-churning customers.

<Table 2: area vs churn

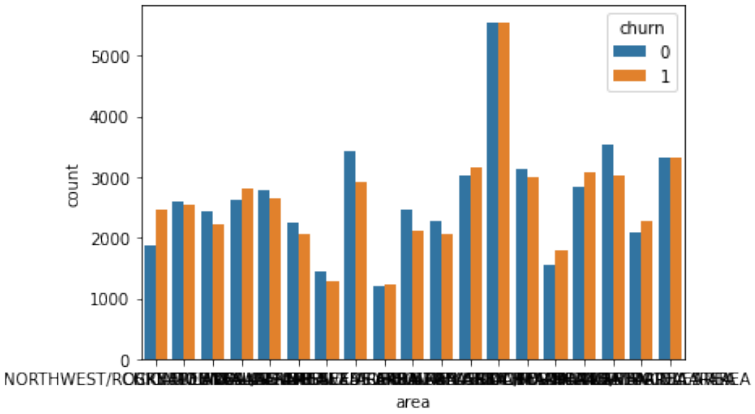

Fig 3 : area vs churn

Similarly, based on the cluster of plots for categorical variables shown below, we can conclude that the relationship between categorical variables is unique, as some variables are approximately equally divided among their categories, while others present a much higher category ratio division, which can challenge the predictive analysis of our model.

All of our plots reveal a unique relationship with churn, and since each variable influences our ability to comprehend the behavior of customers that ultimately leads to churn, we will ensure that all of the data is included in the predictive analysis, even if some data is missing. Furthermore, since we observe a great deal of multicollinearity among the customers, we will use an averaging technique to fill in the gaps whenever necessary.

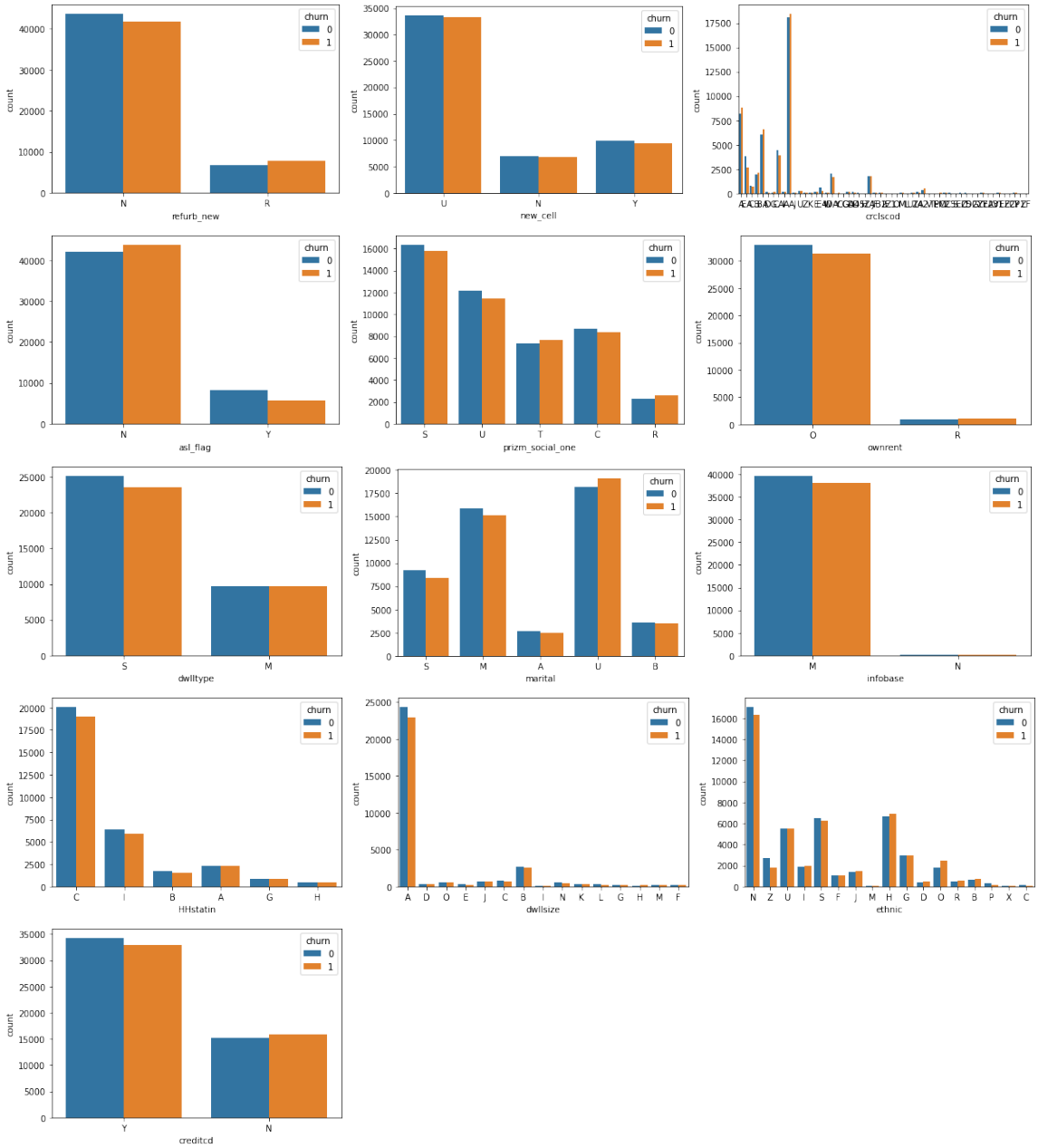

Fig 4: Distribution of categorical variables.

Now that we have completed the exploratory analysis of numerical variables, we can see from Exhibit 3 that the bulk of our numerical features are right-skewed, indicating that there is a probability of outliers in the dataset, which will be addressed in the subsequent report.

In the graph below, we can see that income category 6 consumers made the most calls compared to all other customers and income categories. While categories 5 and 7 have customers with a comparable number of calls, category 2 seems to be similar to categories 5 and 7 but is really an outlier. As the majority of consumers, regardless of their income level, fall within the range of 40,000 to 50,000 total calls, it is safe to assume that call consumption is roughly equivalent across all income categories.

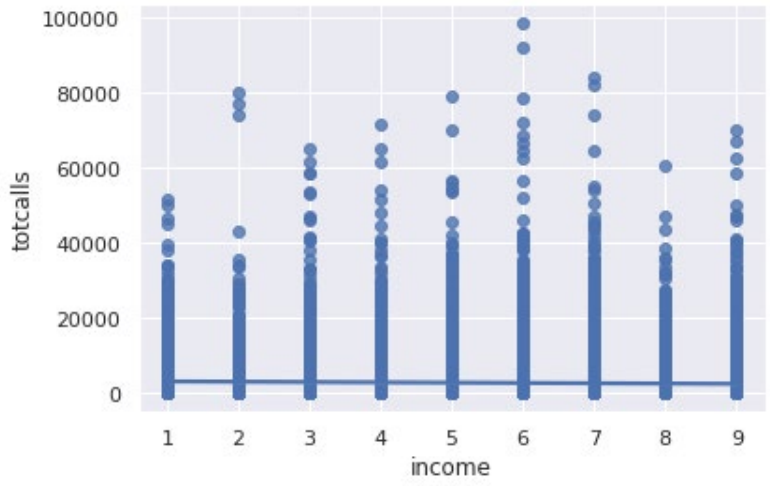

Fig 5: Relation between Totcalls vs income

In the scatterplot on the left, we can see that the average churn across all income groups is between 0.4 and 0.6, i.e. 0.5, indicating that churn is evenly distributed across all income categories. The graph on the right gives more evidence that, despite the fact that the number of income groups differs, the difference in turnover across categories is not particularly large. This indicates that customer attrition across all income brackets is difficult to anticipate.

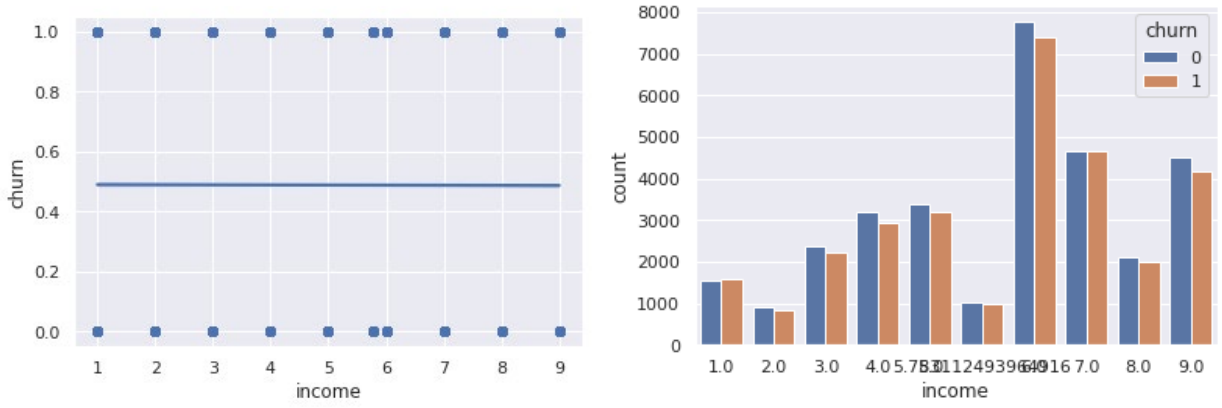

Fig 6: Regression plot and bar polt of churn vs income

The graph below depicts the distribution of income categories across the various regions of the country. For example, S represents the suburbs, where the greatest number of people from various income categories reside, followed by u for urban, t for the town, c for the city, and R for rural, in decreasing

order of each income category count.

Fig 7: Distribution of Prizm_social_one grouped by income category

Data Loading and Pre Processing

Initially, we checked the correlation of all the variables with target variables for missing value treatment and selection of significant features. As customer id is unique to each row and has no bearing on prediction analysis, we removed the Customer ID feature from our dataset. We eliminated a few additional features, including prizm social one, infobase, crclscod, HHStatin, area, ethnic and forgntvl based on the idea that these features may not have a substantial influence on our model, since they may not contribute to the likelihood of customer turnover.

After gaining a thorough grasp of missing values in the dataset, we used the fillna function to replace all missing numerical values with their mean. While dealing with missing values of categorical variables, all null values were replaced with the value 'UNKW'. Only in the instance of the vehicle was the value changed with 0. We used two methods to handle missing data. First, missing value imputation as indicated in the preceding section, and second, eliminating the rows entirely. Rows of columns having less than one percent of missing data and a correlation of low significance have been eliminated. Therefore, after treatment, we utilised the matrix function to produce a plot in which null values are shown horizontally in a Black bar plot as shown in below fig ; this is used to see any missing values in the data set.

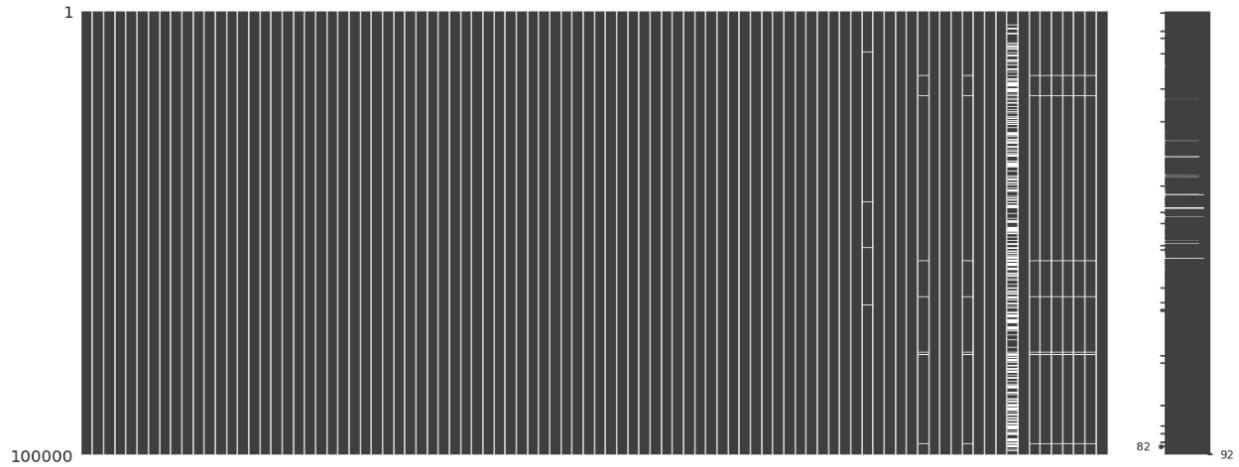

Fig 8: Visualization of null values in NaN matrix

In exhibit 4, it appears that almost all of the features have outliers; for instance, actvsubs and uniqsubs appear to have outliers that fall outside the interquartile range. However, upon examining the dataset and its summary, we discovered that the maximum values in almost all of the features are very similar to the mean or mode, which makes any value which is probable and different from the common value out of IQR.

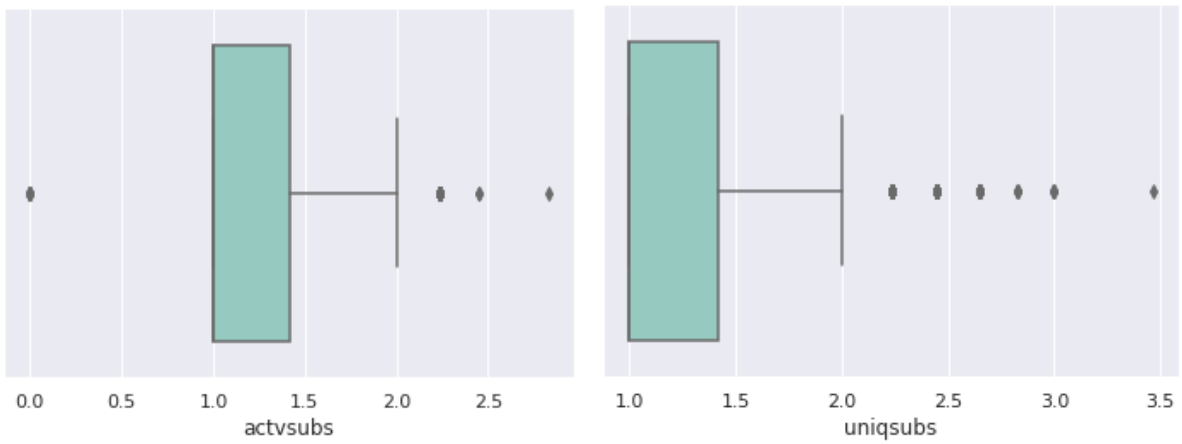

Fig 9: Boxplots of actsubs and uniqsubs

After examining the correlation values of all numerical characteristics, we chose to investigate the correlation of categorical features by expanding all categorical variables using one-hot encoding. It makes our training data more expressive and relevant, and it can be readily rescaled. As numeric values affect our values' probabilities. Before examining the association, categorical variables were converted to numbers using get dummies. Encoding variables having more than two categories with a single value. And after performing one-hot encoding from 92 columns we created 133 columns in total. Then, we evaluated the correlation of each characteristic with churn, sorted them in decreasing order, extracted

the features with the greatest correlation to the goal from the bottom of the list, and displayed the data as shown below:

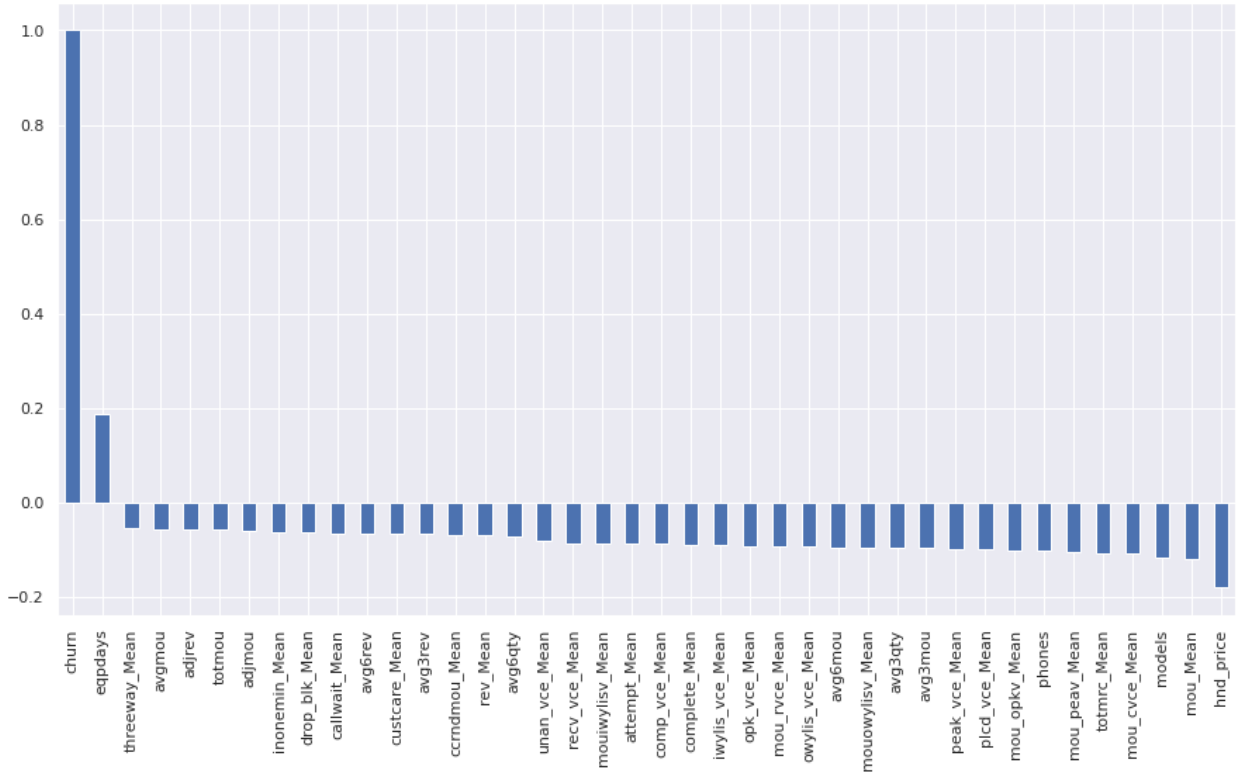

Fig 10: Top 40 correlated features witth churn

From the preceding graphic, we can conclude that even the top 40 highly correlated characteristics are not appropriately linked with the goal feature ie churn.

Models and Methods

Before beginning predictive modelling on the training dataset, we first performed a log modification to address outliers in all features, and then split the dataset into two data frames, with one including all independent features and the other containing all dependent characteristics. Using the train test split() function from the sklearn package, the data set was split 80:20 for modelling and prediction analysis. Now, in order to do predictive analysis, we have selected eight ML models, the first of which is logistic regression, which is required for binary classification issues. To capture the normal distribution of data, we have used the Gaussian Naive Bayes model, as well as the Random forest model, which may be effective for combining many decision trees to provide precise answers. In addition, we employed the XG Boost classifier, which is a machine learning method inside the gradient boosting framework that can handle classification models with a nonlinear connection. We have also used the gradient boosting model and the decision tree model due to their ability to successfully combat bias error and accuracy, respectively. The ADA boost model is used since it is helpful if the model is a poor learner. We have also

employed the ANN model with 4 layers and 100 epochs, however in this model, RELU was used as the activation function to combat non-linearity.

Analysis and Results

In the course of our study, we used a large number of models and analysed the results in order to acquire the greatest possible degree of precision. To do this, we choose the model that helps reduce the frequency of false negative outcomes generated by models in the confusion matrix. The confusion matrix is a table that compares actual vs anticipated numbers as shown in figure below.

		Actual	
		Positive	Negative
Predicted	Positive	TP	FP
	Negative	FN	TN

Fig 11: Confusion matrix

False Positives, which occur when a client who did not churn is incorrectly predicted by our models as having churned, are the major focus of our efforts, and we are striving to reduce their occurrence as much as possible. In addition to accuracy, we are also concentrating on recall, precision, and f1 score.

- Recall: It indicates how many subscribers the algorithm really lost. It aims to establish how many true positives were properly classified.
- Precision: It reveals what proportion of churned subscribers have really left the list.
- F1 score: It is the statistic used to evaluate the performance of algorithms. Due to the unequal classes in the case study, F1-score was suggested.

In the quest of precision, we chose to begin modelling using Gaussian Naive Bayes.

Gaussian Naive Bayes

In our Gaussian model, we received scores of 0.557 for accuracy, 0.534 for precision, 0.637 for recall, and 0.554 for f1 score. The original data were compared to these scores. The relatively low accuracy score suggests that this model is unable to replicate the results it attained with the training dataset when applied to the testing dataset. In addition, after altering the hyperparameter values, we did not see a substantial improvement in the results. The following confusion matrix demonstrates that the

number of false positives is 2156, while the ROC curve plot reveals that the AUC (area under the curve) is 0.578. These two values are shown below. On the basis of the above-mentioned data, we reasoned that other models may provide even better results.

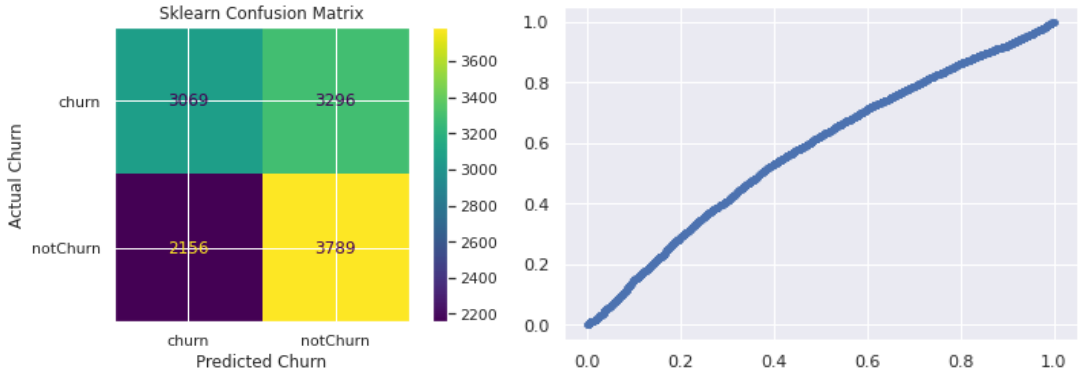

Fig 12: Confusion matrix and ROC curve plot for Gaussian Naive Bayes

Ada Boost

In the Ada model, we received scores of 0.605 for accuracy, 0.593 for precision, 0.584 for recall, and 0.605 for f1 score. There is an improvement in accuracy as compared to the previous model but it is still not significant and we can perform better with other models. The following confusion matrix demonstrates that the number of false positives is 2472, while the ROC curve plot reveals that the AUC (area under the curve) is 0.652.

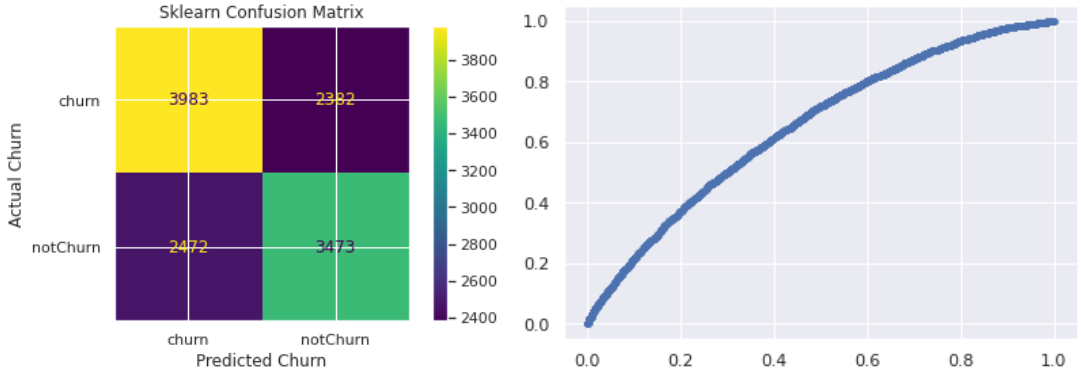

Fig 13: Confusion matrix and ROC curve plot for Ada Boost

Logistic Regression

In our third attempt, we applied logistic regression for the binary classification issue and we received a 0.608 accuracy rate, a 0.598 precision rate, where we received a recall rate of 0.576, and an f1 score of 0.608. after this, we tried hyperparameter tuning with the model and here also we got no improvement in our results. Till here we have improved the accuracy but we will try to test other models for more accuracy. We received 2519 false-positive values in the confusion matrix whereas for the ROC graph the AUC value is 0.649.

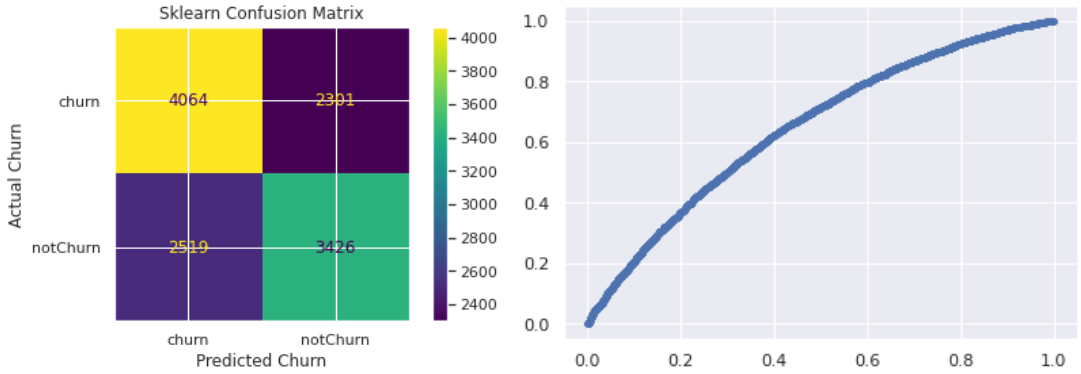

Fig 14: Confusion matrix and ROC curve plot for logistic regression

Gradient Boosting

We tested the Gradiant Boosting Classifier model and obtained the following results: 0.614 for accuracy, 0.599 for precision, 0.605 for recall, and 0.614 for f1 score. When we altered the hyperparameter settings, we did not see a significant increase in the results, despite the fact that this model yields a noticeable performance boost. The accompanying confusion matrix illustrates that there were 2,346 false positives, while the ROC curve plot reveals that the AUC (area under the curve) is 0.664.

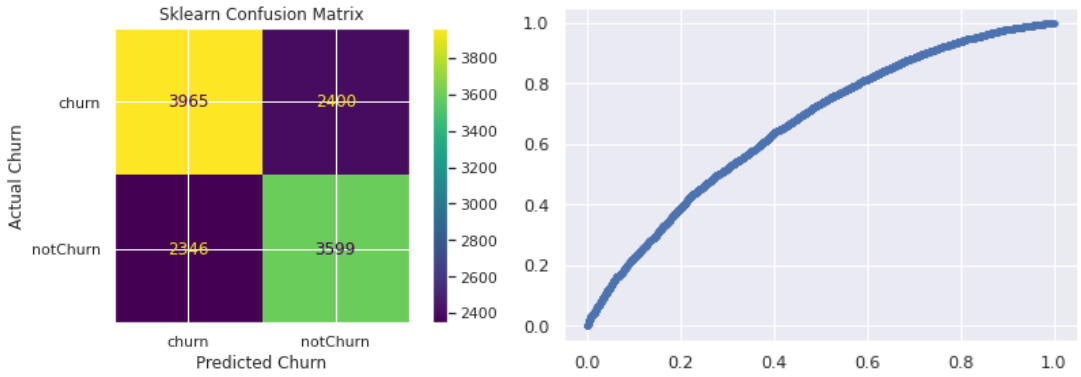

Fig 15: Confusion matrix and ROC curve plot for gradient boosting

XG Boosting

XG Boost model scored 0.614 for accuracy, 0.60 for precision, 0.601 for recall, and 0.614 for f1 score. This model results in no improvement in the performance in comparison to Gradient Boosting, but it's still improved as compared to the rest of the models when we tried altering the hyperparameter values, we did not see a substantial improvement in the results. The following confusion matrix demonstrates that the number of false positives is 2370, while the ROC curve plot reveals that the AUC (area under the curve) is 0.658.

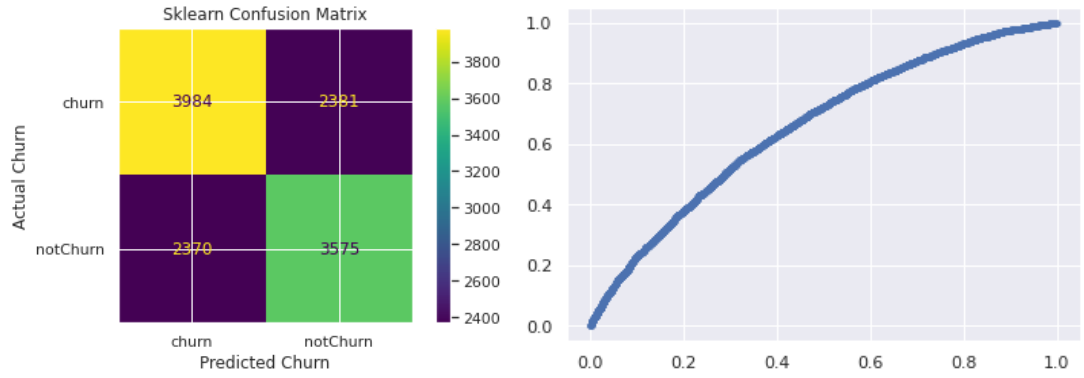

Fig 16: Confusion matrix and ROC curve plot for XG Boosting

Random Forest

The random forest model scored 0.613 for accuracy, 0.603 for precision, 0.577 for recall, and 0.612 for f1 score. This model results in no improvement in the performance in comparison to the previous two models, but when we tried altering the hyperparameter values, we did see some improvement in the results. The following confusion matrix demonstrates that the number of false positives is 2599, while the ROC curve plot reveals that the AUC (area under the curve) is 0.655.

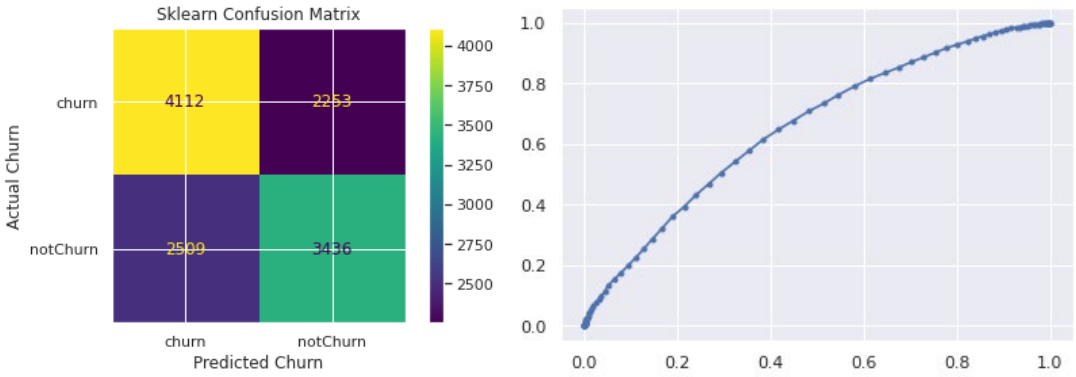

Fig 17: Confusion matrix and ROC curve plot for random forest

We also attempted to use a model called Decision Tree Classification, but it produced unsatisfactory results, so we abandoned that approach. The findings that we obtained are as follows.

accuracy: 0.5446791226645004

precision: 0.5281643472498343

recall: 0.5362489486963835

f1_score: 0.544769144750506

ANN (Artificial Neural Network)

After using all other models for estimating the churn rate, we also attempted the well-known ANN model, which was suggested and used by several researchers. We inserted 64, 256, 512, and 1 neurons in layers 1, 2, 3, and 4, respectively, in the neural network utilised for this model. We added the RELU activation function to the third of four layers since it is a non-linear activation function that does not stimulate all neurons simultaneously. In contrast, the fourth neuron layer features a sigmoid function that is suitable for models in which the output probability must be predicted between two discrete values (0 to 1). While fitting the model, we maintained the ephocs at 100 and followed the loss and accuracy trend for training and testing the data set. This model has overfitted outcomes, as seen by the loss graph, which shows that for the training data set, losses are extremely near to 0 and there is no overlap between testing and training. On the other hand, training accuracy is 1 and testing accuracy is less than 0.6. In this instance, the use of ANN is not a wise choice.

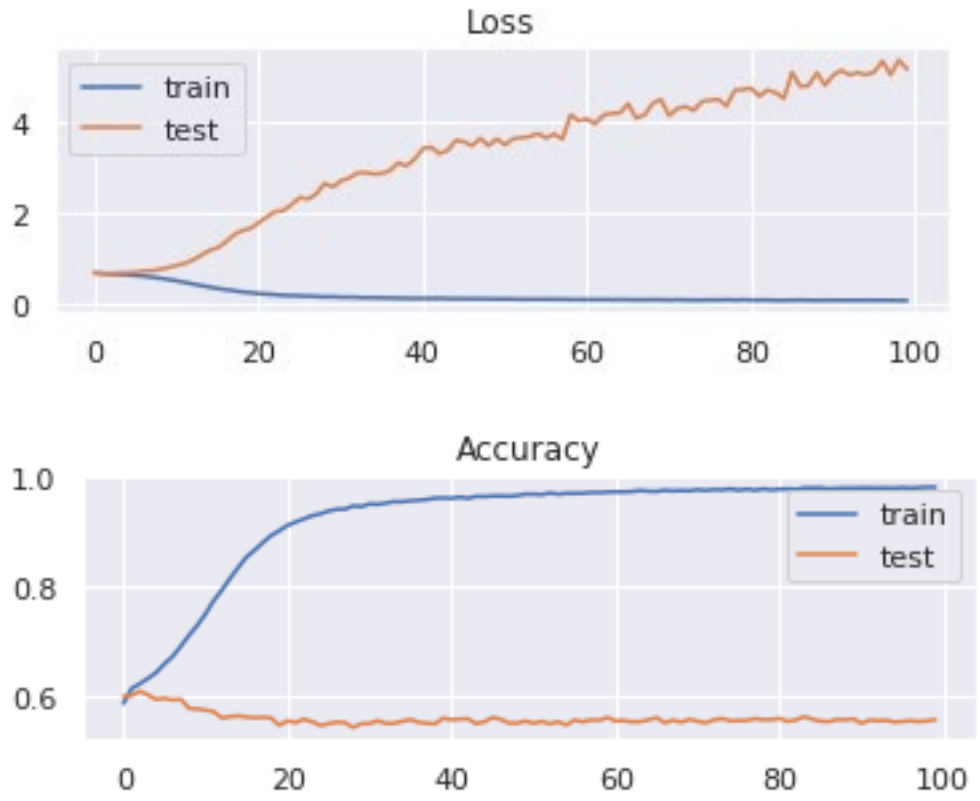

Fig 18: ANN Loss and accuracy graph

Our ML-based modelling Initially, I used the Gaussian Naive Bayes model, which yielded disappointing findings since they imply that the model is not very effective with the data set, resulting in the testing data set having underfitted outcomes. Consequently, we continued with the Ada Boost model, which is a very successful model for tackling classification issues such as churn, and based on the findings, we can conclude that accuracy and other parameters have improved. Following this, we implemented the Logistic Regression model, which did not result in a substantial improvement over the Ada boost model.

Following this, we used Gradient Boosting and XG Boosting, which are regarded as efficient models since they are based on a decision tree structure, which seems to be a more effective approach to manage the data, as is also obvious from the table of summary results below. Then we tried Random Forest, which is also extremely effective at handling classification difficulties and employs a decision tree framework to solve the problem, and as we can see from the results summary table, the results are not as good as those of the first two models but are better than the first three. We also attempted Decision Tree Classification, but the results were inadequate. Due to the intricacy of the data, we opted to employ ANN (Artificial Neural Network) modelling to create a more accurate model, however even with this model, we encountered a great deal of overfitting with the training dataset despite experimenting with various hyperparameter combinations.

Models	Accuracy	Precision	Recall	f1_score	AUC
Gaussian Naive Bayes	0.557	0.534	0.637	0.554	0.578
Ada Boost	0.605	0.593	0.584	0.605	0.652
Logistic Regression	0.608	0.598	0.576	0.608	0.649
Gradient Boosting	0.614	0.599	0.605	0.614	0.664
XG Boosting	0.614	0.6	0.601	0.614	0.658
Random Forest	0.613	0.603	0.577	0.612	0.655
Decision Tree Classification	0.544	0.528	0.536	0.544	0.544
ANN (Artificial Neural Network)	0.557	NA	NA	NA	NA

Table 3: Summary table of model results

The objective of the research was to discover which classification algorithm would be the most successful in predicting customer turnover, given that there are several alternative classification methods. When picking the most effective algorithm, we considered both its performance during the validation process and its precision when generating predictions based on actual test data. The most accurate models were Gradient and XG Boosting, while the least accurate model was the Decision Tree Classification Model. Randomforest provides the maximum degree of precision. With gradient Boosting, recall, accuracy, and overall performance are all improved. As a consequence, we can conclude that Gradient Boosting is the ideal model for resolving the churn problem in the dataset we selected. The ROC is a probability curve, and the AUC is the separability degree, sometimes known as a separability measure. It displays how well the model is able to distinguish between distinct classes. The greater the area under the curve (AUC), the more correctly the model can predict zero classes as zero and one classes as one. Our claim that the Gradient Boosting model is preferable is bolstered by the fact that it was found to have the highest area under the curve (AUC).

Discussion and Conclusion

Due to the amount of data and the non-linear nature of the correlations, we decided to utilise Google Colab, which is a very useful tool, and it made doing the research much easier. Although the predictive model established in this inquiry was able to produce predictions, the accuracy of those forecasts was not one hundred percent. This might have been the situation due to the model's limitations. Our study's

objective was to examine whether or not a certain telecom company's dataset included sufficient information about its customers to forecast the risk of client loss. The noise in the data has reduced the predictive power of the data, despite the fact that our data collection comprises one hundred various elements of a customer, one of which is a target variable that indicates whether or not the consumer is cancelling their account. Throughout the course of our research, we encountered a number of challenges with the dataset owing to a few restrictions, which will be discussed in further depth below.

Limitations:

The target variable had the highest correlation with hnd price and eqpdays (i.e., 0.120929 and 0.124737, respectively), which is normally quite low for prediction purposes. Limitation: Our primary restriction or obstacle was the poor correlation between all of the variables. Even the goal variable, churn, does not correlate significantly with any other variable.

This brings us to our second barrier, the limited availability of resources. Due to the complexity of the relationship between the attributes, the predictive analysis required a substantial amount of computer power, which was scarce.

We need a high degree of accuracy in our forecasts since we lacked a large number of parameters that were highly connected with one another. In order to compile a census of consumer behaviour, the data sets must include additional socioeconomic and demographic characteristics.

Another issue we faced with the data was that the overwhelming majority of characteristics had zero values. Even though we attempted to fix it using the log transformation method, the outcome was not very accurate. The inclusion of zero values in the data gave the impression that the valid values were out of the ordinary, but this was not the case.

The data only contains information from 19 cities, towns, and other localities; hence, it does not accurately depict the country as a whole. It is also possible that the data just relates to a certain section of the United States, such as the southwestern region. To do good predictive analysis, we must, however, cover as much terrain as feasible.

Conclusion :

In our examination of a US telecom company's customer data, we attempted to understand why consumers are quitting the company's service. Our explanation was based on characteristics describing the characteristics of the telecom business and numerous variables deemed crucial when interacting with telecom clients. Even though the data set had 100 distinct qualities, some of those attributes were similar. Despite the relatively low correlation, we used a range of charts and plots to illustrate the relationship between variables during our first inquiry involving the objective variable. There were several difficulties with the data, since some crucial variables had zero values, resulting in a poor correlation with the target variables. In addition, some significant variables had a large number of missing values, which we addressed by imputing the missing values and removing missing value rows of low importance. There were several difficulties with the data, since some crucial variables had zero

values, resulting in a poor correlation with the target variables. Then, after applying a variety of models, we determined that the Gradient Boosting model was the best option since it provided the highest accuracy, area under the curve (AUC), and recall rates. Even though the results were not precisely what we anticipated, we have reason to think that, with enhanced data processing, this model will be able to give us with significant insights and aid us in making crucial business decisions. In other words, the quality of the data will undoubtedly have an influence on the correlation values with the target feature, resulting in a better rate of correct judgement and a larger ability for decision-making.

Future Application:

Because we have previously addressed the utility of this model, despite its numerous flaws, it may be used to develop methods for reducing customer turnover and fostering customer loyalty over the long term. As previously noted, this may be performed by targeting individual clients that meet certain criteria. This model may be used to analyse churn in other regions of the United States, given that the data quality is either already excellent or can be enhanced via processing. In addition, updated models may provide us with more precise findings, despite the complexity of the data.

References

Alwis, P.K.D.N.M., Kumara, B.T.G.S. and Hapuarachchi, H.A.C.S., 2018. Customer Churn Analysis and Prediction in Telecommunication for Decision Making.

Dahiya, K. and Talwar, K., 2015. Customer churn prediction in telecommunication industries using data mining techniques-a review. Int. J. Adv. Res. Comput. Sci. Softw. Eng, 5(4), pp.417-433.

He, Y., He, Z. and Zhang, D., 2009, August. A study on prediction of customer churn in fixed communication network based on data mining. In 2009 sixth international conference on fuzzy systems and knowledge discovery (Vol. 1, pp. 92-94). IEEE.

Huang, Y., Zhu, F., Yuan, M., Deng, K., Li, Y., Ni, B., Dai, W., Yang, Q. and Zeng, J., 2015, May. Telco churn prediction with big data. In Proceedings of the 2015 ACM SIGMOD international conference on management of data (pp. 607-618).

Idris, A., Khan, A. and Lee, Y.S., 2012, October. Genetic programming and adaboosting based churn prediction for telecom. In 2012 IEEE international conference on Systems, Man, and Cybernetics (SMC) (pp. 1328-1332). IEEE.

Nigam, B., Dugar, H. and Niranjanamurthy, M., 2019. Effectual predicting telecom customer churn using deep neural network. Int J Eng Adv Technol (IJEAT), 8(5).

Tables and Figures Index

Appendix

Exhibit 1

S.No	Columns	Description
1	rev_Mean	Mean monthly revenue (charge amount)
2	mou_Mean	Mean number of monthly minutes of use
3	totmrc_Mean	Mean total monthly recurring charge
4	da_Mean	Mean number of directory assisted calls
5	ovrmou_Mean	Mean overage minutes of use
6	ovrrev_Mean	Mean overage revenue
7	vceovr_Mean	Mean revenue of voice overage
8	datovr_Mean	Mean revenue of data overage
9	roam_Mean	Mean number of roaming calls
10	change_mou	Percentage change in monthly minutes of use vs previous three month average
11	change_rev	Percentage change in monthly revenue vs previous three month average
12	drop_vce_Mean	Mean number of dropped (failed) voice calls
13	drop_dat_Mean	Mean number of dropped (failed) data calls
14	blck_vce_Mean	Mean number of blocked (failed) voice calls
15	blck_dat_Mean	Mean number of blocked (failed) data calls
16	unan_vce_Mean	Mean number of unanswered voice calls
17	unan_dat_Mean	Mean number of unanswered data calls
18	plcd_vce_Mean	Mean number of attempted voice calls placed
19	plcd_dat_Mean	Mean number of attempted data calls placed
20	recv_vce_Mean	Mean number of received voice calls
21	recv_sms_Mean	Mean number of messages received
22	comp_vce_Mean	Mean number of completed voice calls
23	comp_dat_Mean	Mean number of completed data calls
24	custcare_Mean	Mean number of customer care calls
25	ccrndmou_Mean	Mean rounded minutes of use of customer care calls
26	cc_mou_Mean	Mean unrounded minutes of use of customer care (see CUSTCARE_MEAN) calls
27	inonemin_Mean	Mean number of inbound calls less than one minute
28	threeway_Mean	Mean number of three way calls
29	mou_cvce_Mean	Mean unrounded minutes of use of completed voice calls
30	mou_cdat_Mean	Mean unrounded minutes of use of completed data calls
31	mou_rvce_Mean	Mean unrounded minutes of use of received voice calls
32	owylis_vce_Mean	Mean number of outbound wireless to wireless voice calls
33	mouowylisv_Mean	Mean unrounded minutes of use of outbound wireless to wireless voice calls
34	iwylis_vce_Mean	inbound wireless to wireless voice calls mean
35	mouiwylisv_Mean	Mean unrounded minutes of use of inbound wireless to wireless voice calls
36	peak_vce_Mean	Mean number of inbound and outbound peak voice calls
37	peak_dat_Mean	Mean number of peak data calls
38	mou_peav_Mean	Mean unrounded minutes of use of peak voice calls
39	mou_pead_Mean	Mean unrounded minutes of use of peak data calls
40	opk_vce_Mean	Mean number of off-peak voice calls
41	opk_dat_Mean	Mean number of off-peak data calls
42	mou_opkv_Mean	Mean unrounded minutes of use of off-peak voice calls
43	mou_opkd_Mean	Mean unrounded minutes of use of off-peak data calls
44	drop_blk_Mean	Mean number of dropped or blocked calls
45	attempt_Mean	Mean number of attempted calls
46	complete_Mean	Mean number of completed calls
47	callfwdv_Mean	Mean number of call forwarding calls
48	callwait_Mean	Mean number of call waiting calls
50	months	Total number of months in service
51	uniqsubs	Number of unique subscribers in the household
52	actvsubs	Number of active subscribers in household
53	new_cell	New cell phone user
54	crclscod	Credit class code
55	asl_flag	Account spending limit
56	totcalls	Total number of calls over the life of the customer
57	totmou	Total minutes of use over the life of the cus
58	totrev	Total revenue
59	adjrev	Billing adjusted total revenue over the life of the customer
60	adjmou	Billing adjusted total minutes of use over the life of the customer
61	adjqty	Billing adjusted total number of calls over the life of the customer
62	avgrev	Average monthly revenue over the life of the customer
63	avgmou	Average monthly minutes of use over the life of the customer
64	avgqty	Average monthly number of calls over the life of the customer
65	avg3mou	Average monthly minutes of use over the previous three months
66	avg3qty	Average monthly number of calls over the previous three months
67	avg3rev	Average monthly revenue over the previous three months
68	avg6mou	Average monthly minutes of use over the previous six months
69	avg6qty	Average monthly number of calls over the previous six months
70	avg6rev	Average monthly revenue over the previous six months
71	prizm_social_one	Social group letter only
72	area	Geogrpahic area
73	dualband	Dualband
74	refurb_new	Handset: refurbished or new
75	hnd_price	Current handset price
76	phones	Number of handsets issued
77	models	Number of models issued
78	hnd_webcap	Handset web capability
79	truck	Truck indicator
80	rv	RV indicator
81	ownrent	Home owner/renter status
82	lor	Length of residence
83	dwlltype	Dwelling Unit type
84	marital	Marital Status
85	adults	Number of adults in household
86	infobase	InfoBase match
87	income	Estimated income
88	numbcars	Known number of vehicles
89	HHstatin	Premier household status indicator
90	dwllsize	Dwelling size
91	forgntvl	Foreign travel dummy variable
92	ethnic	Ethnicity roll-up code
93	kid0_2	Child 0 - 2 years of age in household
94	kid3_5	Child 3 - 5 years of age in household
95	kid6_10	Child 6 - 10 years of age in household
96	kid11_15	Child 11 - 15 years of age in household
97	kid16_17	Child 16 - 17 years of age in household
98	creditcd	Credit card indicator
99	eqpdays	Number of days (age) of current equipment

Exhibit 2

	No. missing values	% of missing data
numbcars	9565	42.080950
dwllsize	6745	29.674439
HHstatin	6584	28.966124
ownrent	5649	24.852618
dwlltype	5337	23.479982
hnd_webcap	5147	22.644083
lor	4927	21.676199
income	4037	17.760669
adults	3610	15.882094
infobase	3443	15.147382
prizm_social_one	1408	6.194457
kid0_2	331	1.456225
kid3_5	331	1.456225
truck	331	1.456225
forgntvl	331	1.456225
ethnic	331	1.456225
kid16_17	331	1.456225
kid6_10	331	1.456225
kid11_15	331	1.456225
marital	331	1.456225
creditcd	331	1.456225

Exhibit 3

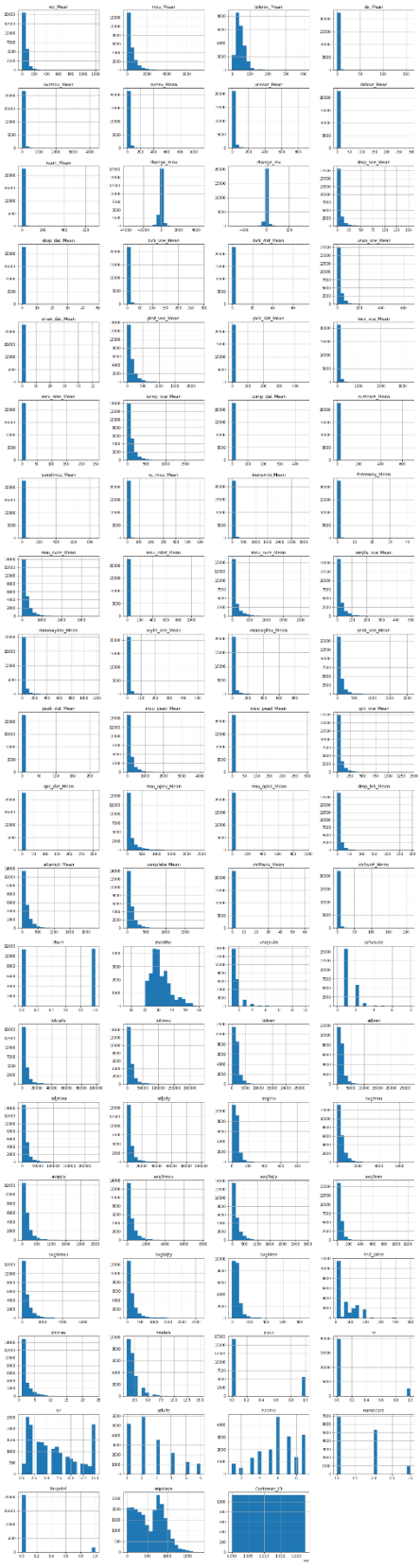

Exhibit 4