TELECOM CHURN PREDICTION

ISM6136.001F22.92628

DATA MINING

TEAM

JYOTHI VENKATA NAGA SAI KUMAR MATURI
DHANUNJAN REDDY RAGHURAM
RANJITH KUMAR KANUMILLI
SAMPREETH MATURI
AKSHAY RAMESH

TELECOM CHURN PREDICTION

Abstract:

When a client base of a company stops doing business with the company, it is referred to as the customer churn rate or attrition rate. It is usually that chunk of the user base who terminate their service contracts with the company before a stipulated time frame. For an organization to earn a profit, its overall growth rate must be higher than its churn rate. In the telecommunications business, the churn rate is a valuable metric. It includes satellite or cable television operators, Internet service providers, and phone services.
Because most consumers have many choices, the churn rate can assist a business in analyzing how it compares to its counterparts in the marketplace. The best way for a company to prevent customer turnover, though, is to have a thorough understanding of its clientele. By employing Data Mining and Data analytics to obtain insights into consumers' expectations and problems, businesses may anticipate them. This enables them to meet their demands and maintain the security and strength of their company. This involves spotting customers who are likely to leave and working to increase their pleasure. Churn analytics is useful for foretelling customer churn and figuring out its underlying causes. Churn analytics is valuable for predicting customer churn and identifying the root causes. The number of consumers who discontinue a product or service during a specific period is called churn.
In our project, we are applying predictive models to forecast attrition on a per-customer basis and adopt ways to reduce it, like providing better discounts, special offers, or other rewards to keep customers. Customer churn analysis is a typical classification task in the supervised learning domain.

Problem Statement:

The major problem faced by the most Telecommunications business is their withdrawing customers who drop their services provided by the company and turn their back on them to shift to another company. The churn rate can be a deciding element in assessing an organization's competitiveness when compared to its competitors in the market because it determines its loyal client base, which determines whether the company will continue to do business in the long run. It is quite simple for clients to switch to another firm if they are no longer satisfied with the service supplied because the services provided by the Telecommunication businesses, like internet service providers or phone service providers, are practically identical to each other in the market.
The organizations must employ superior analytics to anticipate client requests and concerns, satisfy their expectations, gather insights from them, and use those insights to lower the churn rate to keep their business operating and secure. In this project, we hope to use predictive models to predict customer attrition and to offer them greater discounts and special deals to keep them with the business. This would reduce customer churn and help them retain a secure customer base for their business to go on.

Methodology:

Our problem statement deals with predicting whether a person will be likely to get a heart disease or not, we use classification models on our data. We divided our dataset into train and test splits and trained our training data using classification models to find the patterns in the data for determining the likelihood of heart disease prediction and tested our model on the test dataset and get the relevant evaluation metrics to assess the model's performance.

Evaluation Metrics Used:

Confusion Matrix: Confusion Matrix shows us the number of data instances which are classified into Positive class and Negative class from the Actual Positive and Negative classes. The instance which is from positive class classified as positive class it is termed as True Positive. The instance which is from negative class classified as negative class it is termed as True Negative. The instance which is from positive class classified as negative class it is termed as False Negative. The instance which is from negative class classified as positive class it is termed as False Positive.

Precision and Recall: Precision: Out of all the predicted positive classes how many of them are positive. It is the ratio of True Positives (TP) to the Predicted Positive Classes (TP + FP). Recall: Out of all the actual positive classes how many are predicted positive. It is the ratio of True Positives (TP) to the Actual Positive Classes (TP+FN) In all our models in this project we tried to decrease the false negatives cases, which means we are trying to increase the Recall Score.

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive} \qquad \text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

Accuracy Score: Accuracy score indicates the extent of accuracy for the selected model after employing the training and test data sets.

Area Under Curve (AUC): The Area Under the Curve (AUC) is a summary of the ROC curve that measures a classifier's capacity to differentiate among classes. The AUC indicates how well the model distinguishes between positive and negative classes. The greater the AUC, the better it is.

ROC Curve: The trade-off between the true positive rate and the false positive rate for a predictive model utilizing distinctive probability thresholds is summarized by ROC Curves.

F1 Score: The weighted average of Precision and Recall is the F1 Score. As a result, this score considers both false positives and false negatives. It builds harmonic mean of precision & recall and hence calculates the compromise between them.

Dataset Details:

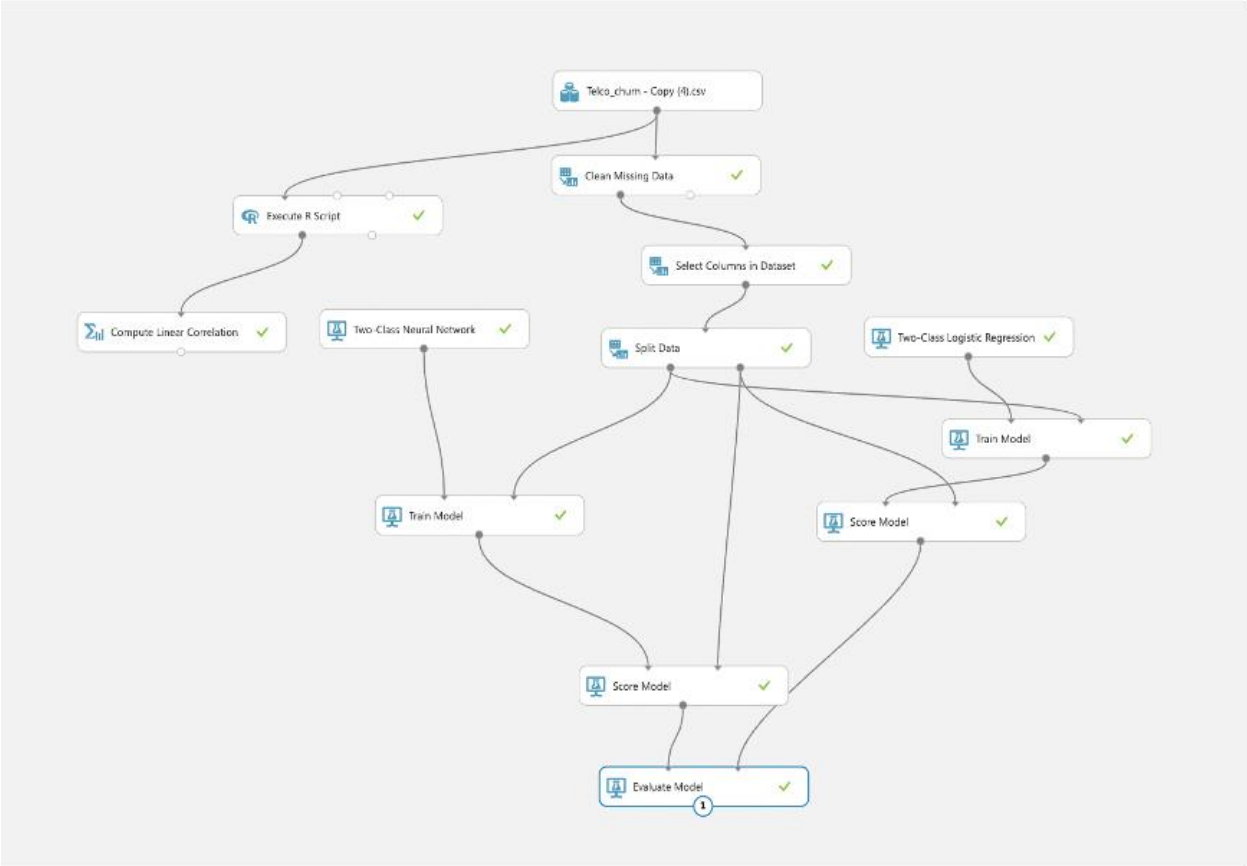
Our team used a dataset from Kaggle called "Telco Customer Churn" to complete this project. Following a closer examination of the dataset, the following traits and parameters can be explained:

1) **Customer ID:** This is a one-of-a-kind identifier sequence used to denote individual dataset records.
2) **Partner:** This column gives information about the customer partner.
3) **Gender:** This column gives information about the gender of the customer in that row (male or female).
4) **Senior Citizen:** Tell us whether the customer is a senior citizen.
5) **Dependents:** Tells whether that customer has any dependents.
6) **Tenure:** It tells us about the total duration for which the customer used the company's service.
7) **Phone Service:** Tell us whether the customer has subscribed to the company's phone service.
8) **Multiple Lines:** This tells us whether the customer has subscribed to multiple lines from the service provider.
9) **Internet Service:** This tells whether the user has subscribed to the service provider's internet services.
10) **Online Security:** This tells us whether the user has subscribed to the provider's online security feature.
11) **Online Backup:** This tells us whether the user can access the Online Backup feature.
12) **Device Protection:** This tells us whether the user has device protection.
13) **Tech Support:** This tells us whether the user utilizes the tech support functionality.
14) **Streaming TV:** Tells us whether the customer has subscribed to the provider's TV Streaming services.
15) **Streaming Movies:** Tells us whether the customer has subscribed to the provider's Movie Streaming services.
16) **Contract:** Mentions the contract term for that customer.
17) **Paperless Billing:** Mentions whether the customer has opted for Paperless Billing.
18) **Payment Method:** It will specify the type of billing/payment method employed by the user/customer to make service-related payments.
19) **Monthly Charges:** Tells us the monthly charges that were charged to the user.
20) **Total Charges:** This tells us about the total charges to date that have been charged to the user.
21) **Churn:** This will be our TARGET/DEPENDENT VARIABLE. Tells us whether the user will churn or not.

Dataset Snippet:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	customerI	gender	SeniorCitiz	Partner	Dependen	tenure	PhoneServ	MultipleLir	InternetSe	OnlineSecu	OnlineBac	DevicePro	TechSuppc	Streaming	Streaming	Contract	PaperlessB	PaymentM	MonthlyCh	TotalChar	Churn
2	7590-VHVI	Female	0	Yes	No	1	No	No phone	DSL	No	Yes	No	No	No	No	Month-to-	Yes	Electronic	29.85	29.85	No
3	5575-GNV	Male	0	No	No	34	Yes	No	DSL	Yes	No	Yes	No	No	No	One year	No	Mailed che	56.95	1889.5	No
4	3668-QPYI	Male	0	No	No	2	Yes	No	DSL	Yes	Yes	No	No	No	No	Month-to-	Yes	Mailed che	53.85	108.15	Yes
5	7795-CFO	Male	0	No	No	45	No	No phone	DSL	Yes	No	Yes	Yes	No	No	One year	No	Bank trans	42.3	1840.75	No
6	9237-HQIT	Female	0	No	No	2	Yes	No	Fiber optic	No	No	No	No	No	No	Month-to-	Yes	Electronic	70.7	151.65	Yes
7	9305-CDS	Female	0	No	No	8	Yes	Yes	Fiber optic	No	No	Yes	No	Yes	Yes	Month-to-	Yes	Electronic	99.65	820.5	Yes
8	1452-KIOV	Male	0	No	Yes	22	Yes	Yes	Fiber optic	No	Yes	No	No	Yes	No	Month-to-	Yes	Credit carc	89.1	1949.4	No
9	6713-OKO	Female	0	No	No	10	No	No phone	DSL	Yes	No	No	No	No	No	Month-to-	No	Mailed che	29.75	301.9	No
10	7892-POO	Female	0	Yes	No	28	Yes	Yes	Fiber optic	No	No	Yes	Yes	Yes	Yes	Month-to-	Yes	Electronic	104.8	3046.05	Yes
11	6388-TABC	Male	0	No	Yes	62	Yes	No	DSL	Yes	Yes	No	No	No	No	One year	No	Bank trans	56.15	3487.95	No
12	9763-GRS	Male	0	Yes	Yes	13	Yes	No	DSL	Yes	No	No	No	No	No	Month-to-	Yes	Mailed che	49.95	587.45	No
13	7469-LKBC	Male	0	No	No	16	Yes	No	No	No interne	No interne	No interne	No interne	No interne	No interne	Two year	No	Credit carc	18.95	326.8	No
14	8091-TTV	Male	0	Yes	No	58	Yes	Yes	Fiber optic	No	No	Yes	No	Yes	Yes	One year	No	Credit carc	100.35	5681.1	No
15	0280-XJGE	Male	0	No	No	49	Yes	Yes	Fiber optic	No	Yes	Yes	No	Yes	Yes	Month-to-	Yes	Bank trans	103.7	5036.3	Yes
16	5129-JLPIS	Male	0	No	No	25	Yes	No	Fiber optic	Yes	No	Yes	Yes	Yes	Yes	Month-to-	Yes	Electronic	105.5	2686.05	No
17	3655-SNQ	Female	0	Yes	Yes	69	Yes	Yes	Fiber optic	Yes	Yes	Yes	Yes	Yes	Yes	Two year	No	Credit carc	113.25	7895.15	No
18	8191-XWS	Female	0	No	No	52	Yes	No	No	No interne	No interne	No interne	No interne	No interne	No interne	One year	No	Mailed che	20.65	1022.95	No
19	9959-WOF	Male	0	No	Yes	71	Yes	Yes	Fiber optic	Yes	No	Yes	No	Yes	Yes	Two year	No	Bank trans	106.7	7382.25	No
20	4190-MFLI	Female	0	Yes	Yes	10	Yes	No	DSL	No	No	Yes	Yes	No	No	Month-to-	No	Credit carc	55.2	528.35	Yes
21	4183-MYF	Female	0	No	No	21	Yes	No	Fiber optic	No	Yes	Yes	No	No	Yes	Month-to-	Yes	Electronic	90.05	1862.9	No
22	8779-QRD	Male	1	No	No	1	No	No phone	DSL	No	No	Yes	No	No	Yes	Month-to-	Yes	Electronic	39.65	39.65	Yes
23	1680-VDC	Male	0	Yes	No	12	Yes	No	No	No interne	No interne	No interne	No interne	No interne	No interne	One year	No	Bank trans	19.8	202.25	No
24	1066-JKSG	Male	0	No	No	1	Yes	No	No	No interne	No interne	No interne	No interne	No interne	No interne	Month-to-	No	Mailed che	20.15	20.15	Yes
25	3638-WEA	Female	0	Yes	No	58	Yes	Yes	DSL	No	Yes	No	Yes	No	No	Two year	Yes	Credit carc	59.9	3505.1	No

Model Building:

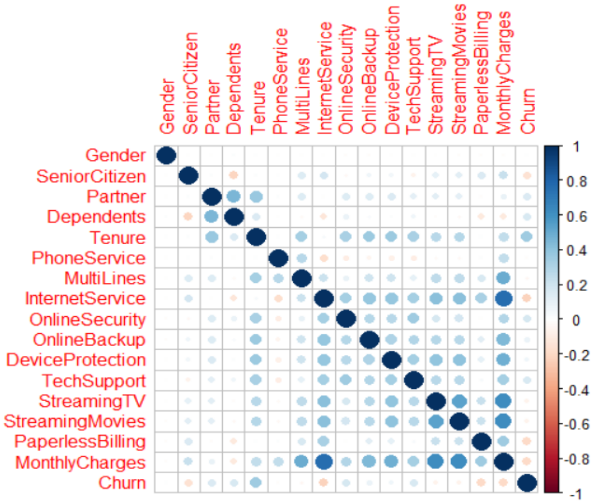


Data Preprocessing:

Cleaning missing data: The first step is to clean the missing data, we are missing data in **total charges** column, we replace it with the mean of the entire column.

Select columns: From the 21 columns in the given dataset, we can see that not all the columns contribute much to the churn. We found 15 out of 21 columns which are contributing more to the churn using the correlation matrix in R.

Correlation Plot we got using R: In this the blue getting darker represents more correlation between the variables.



Split data: We have split the data into 70 and 30 percent with giving 70 to the train and 30 to the test.

Predictive model: In this predictive model, we have employed 4 algorithms/models and compared how well they fared. The input dataset was initially imported, and its significant columns were selected for further modelling purpose. Next, the data is split into training and test data. We split 70% of the data for training purpose and remaining 30% for testing purpose.

Further the following models are used for our purpose:

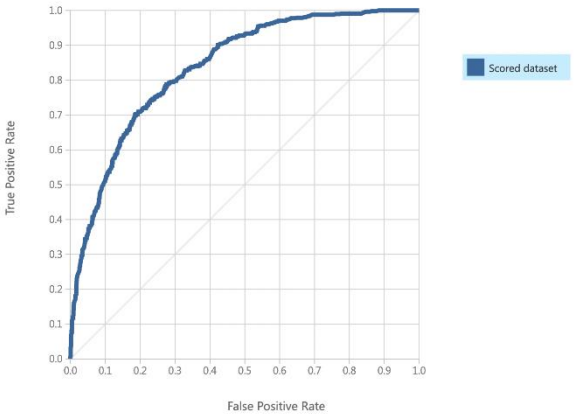
1) **Two-class Neural Network:** A simple neural network is used to compare the two methods, even though the data set is small and that neural networks normally need a lot of training data to be reliable predictors.

2) **Two-class Logistic Regression:** A well-known statistical technique for estimating the likelihood of a particular outcome, logistic regression is particularly helpful for categorization issues.

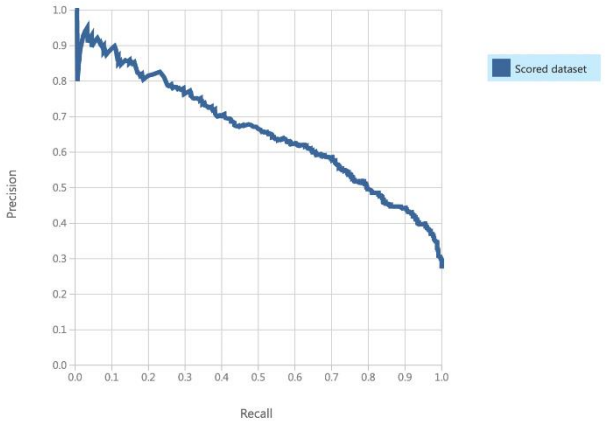
Model Outcomes:

1) Two-class Logistic Regression:

ROC Curve



Precision/Recall Curve

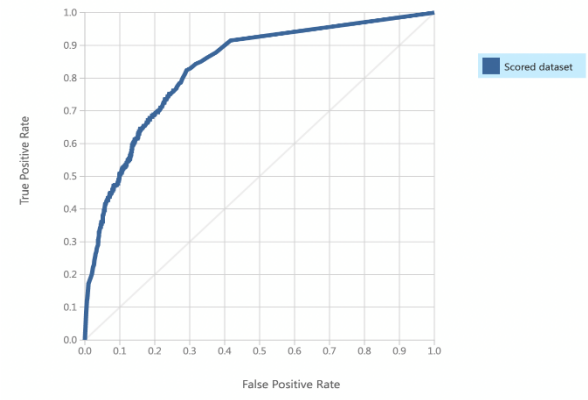


True Positive	False Negative	Accuracy	Precision
311	264	0.794	0.644
False Positive	**True Negative**	**Recall**	**F1 Score**
172	1366	0.541	0.588
Positive Label	**Negative Label**		
Yes	No		

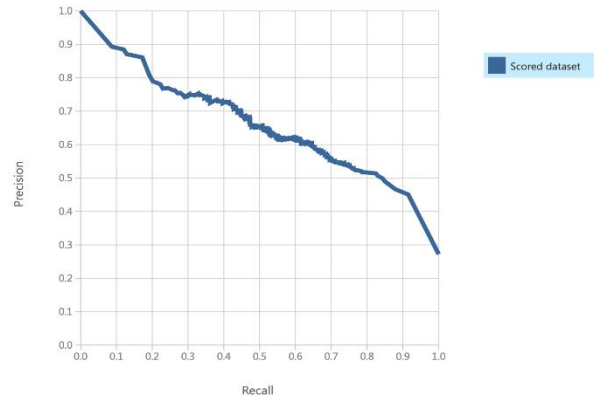
Threshold 0.5 — AUC 0.838

2) Two-class Neural Network:

ROC Curve



Precision/Recall Curve

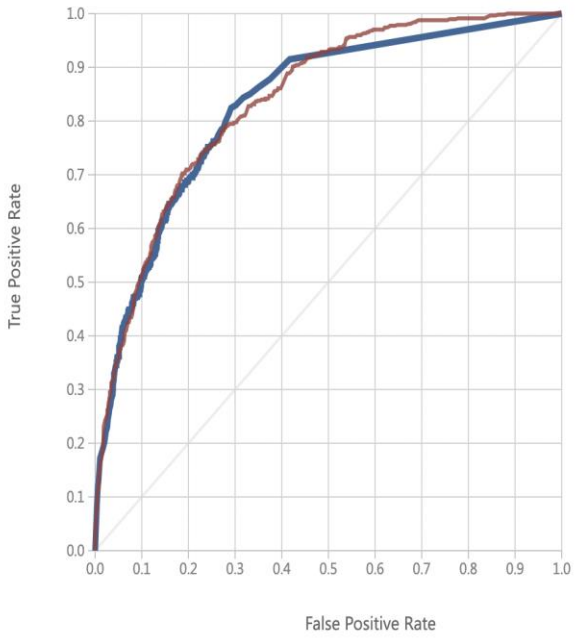


True Positive	False Negative	Accuracy	Precision
314	261	0.788	0.625
False Positive	**True Negative**	**Recall**	**F1 Score**
188	1350	0.546	0.583
Positive Label	**Negative Label**		
Yes	No		

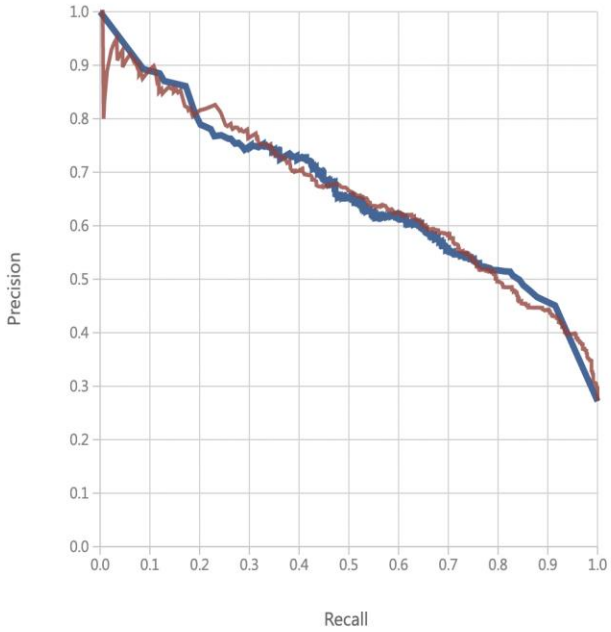
Threshold 0.5 — AUC 0.833

Comparison:

ROC Curve



Precision/Recall Curve



In the above graphs the red curve represents the two-class logistic regression, and the blue curve represents the two-class neural network. Using the above graphs, we've made the following evaluation and conclusions.

Model evaluation:

Based on the above results we can conclude from the data that two class logistic regression outperformed the two-class neural network. We achieved a 79% accuracy in logistic regression using FP & FN values of 264 and 172, respectively.

Conclusion:

For subscription-based businesses, the churn rate is a key metric. The capacity to identify consumers who are dissatisfied with offered solutions allows organizations to gain insight about product or price plan flaws, operational challenges, as well as client preferences, to prevent churn.
Our Machine Learning model may be appropriately trained to obtain higher accuracies if we used them. To be able to identify prospective client scenarios where churn can be avoided and consumers satisfied, a high level of accuracy is required.