Churn Prediction | Industry: Telecom

Author: Jatin Arora | Tools Used: Tableau for Data Analysis and Jupyter notebook for Python

1. **Problem Introduction**
   Telecom companies spend hundreds of dollars to acquire a new customer and when that customer leaves, the company not only loses the future revenue from that customer but also the resources spend to acquire that customer. Churn erodes profitability.

2. **Business Objective**
   The business aims to identify customers which are about to churn from their business and identify the factors which are leading to this loss of business.

3. **Science Objective**
   Science objective is to develop a machine learning model that is able to identify the patterns from the data set and is able to identify the customers which are about to churn from the network. Besides this, the model should be able to explain the factors which are responsible for this, and should have some threshold accuracy

4. **Hidden Questions**
   Devise a strategy to:
   - Increase the retention rate
   - Feature engineer different other features apart from the given data
   - Bring insights from the given data

5. **Problem Approach**
   - Data Import
   - Data Analysis – EDA
   - Data Preparation
   - Preliminary Analysis
   - Univariate Analysis
   - Multivariate Analysis
   - Feature Engineering
   - Modelling -Mental Model Preparation
   - Unsupervised Approach
   - Supervised Approach
   - Validation & Model Selection

6. **Approaches in detail:**

Stage 1: Current Data and definitions

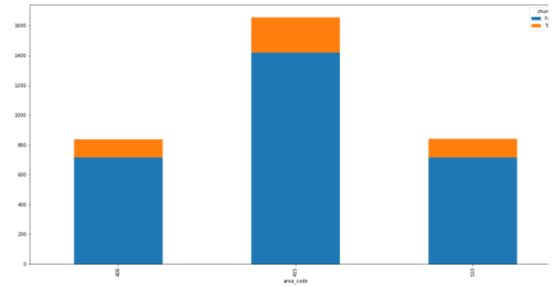
Data Size: 3333 rows/unique customers

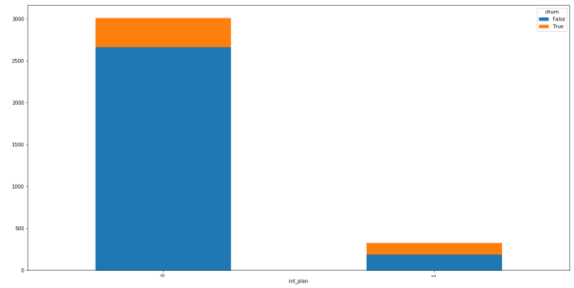
Data Schema:

	count	mean	std	min	0.25	0.5	0.75	max
acc_len	3333	101.0648065	39.82210593	1	74	101	127	243
area_code	3333	437.1824182	42.37129049	408	408	415	510	510
int_plan	3333	0.096909691	0.295879145	0	0	0	0	1
voice_mail_plan	3333	0.276627663	0.44739787	0	0	0	1	1
num_vmail_msg	3333	8.099009901	13.68836537	0	0	0	20	51
total_day_minutes	3333	179.7750975	54.4673892	0	143.7	179.4	216.4	350.8
total_day_calls	3333	100.4356436	20.06908421	0	87	101	114	165
total_day_charge	3333	30.56230723	9.259434554	0	24.43	30.5	36.79	59.64
total_eve_min	3333	200.980348	50.71384443	0	166.6	201.4	235.3	363.7
total_eve_calls	3333	100.1143114	19.92262529	0	87	100	114	170
total_eve_charge	3333	17.08354035	4.310667643	0	14.16	17.12	20	30.91
total_night_min	3333	200.8720372	50.57384701	23.2	167	201.2	235.3	395
total_night_calls	3333	100.1077108	19.56860935	33	87	100	113	175
total_night_charge	3333	9.039324932	2.275872838	1.04	7.52	9.05	10.59	17.77
total_int_min	3333	10.23729373	2.791839548	0	8.5	10.3	12.1	20
total_int_calls	3333	4.479447945	2.461214271	0	3	4	6	20
total_int_charge	3333	2.764581458	0.753772613	0	2.3	2.78	3.27	5.4
customer_service_calls	3333	1.562856286	1.315491045	0	1	1	2	9

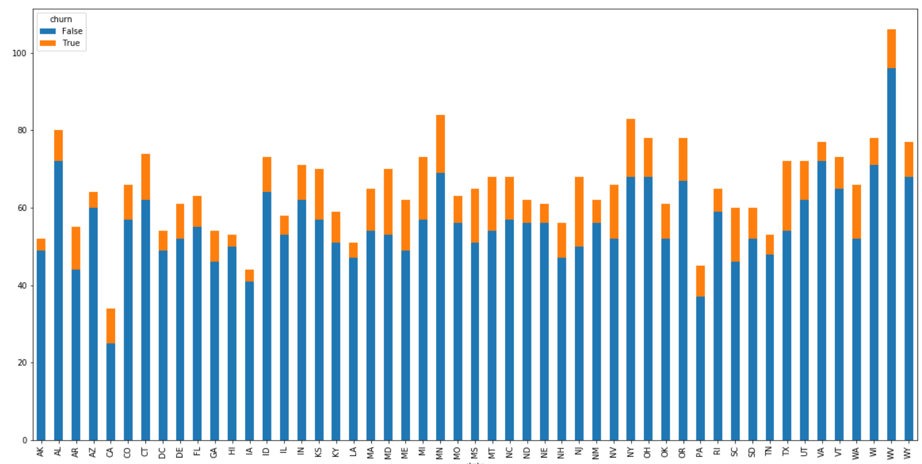
Stage 2: Analysis | Inferences | Visualizations

1. Churn Customer Distribution Across Features: Churn Customers in the different Area Code(Left) and Churn customers in the International Plan

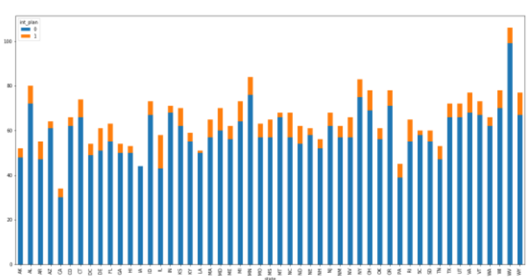


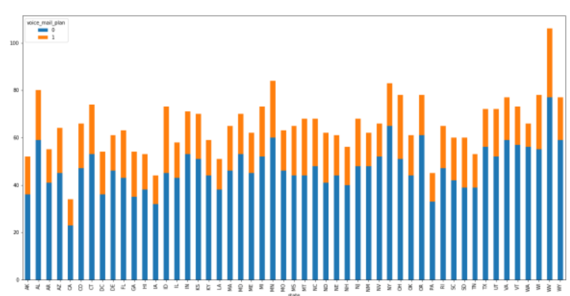


2. Churn Customers in the different State: Certain
   states like MO, MT, NM TX, MI, etc have higher churn customers.

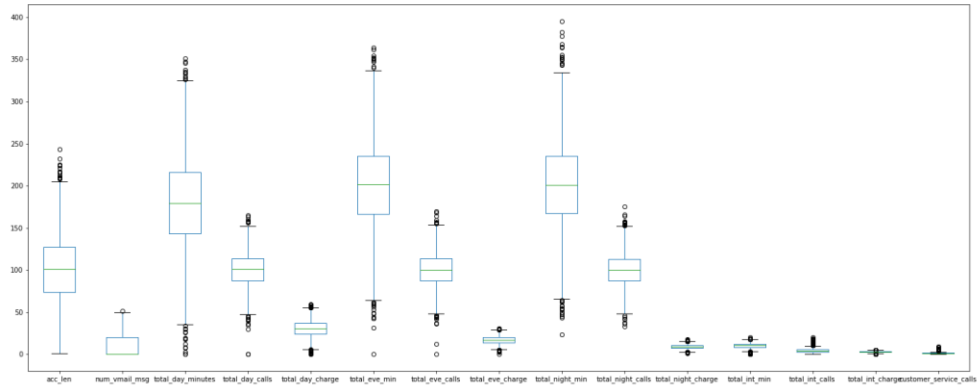


3. Correlation and distribution Plot Appendix
4. Customers who have availed the international (left) and voice mail (right) plan in the different states





5. Distribution of Different features with boxplots

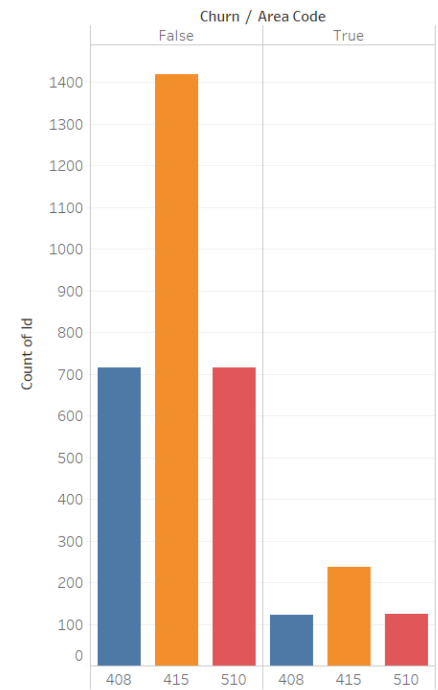


6. There are three area codes where the telecom operator serves:

- 408: Total Customers - 838
- 415: Total Customers - 1655
- 510: Total Customers – 840

Note: The Graph adjacent represents customer churn by Area code

Inference 1: As 415 is a high customer density area the churn rate is also high, perhaps the company would like to arrest the customer churn starting from 415



7. Premium Customer: Those customers which are using both Voice Mail Plan and International Plan. These customers are using two services from the company and are revenue generating than other customers. These customers should at no cost be churned as these are profitable entities.

Below graph shows that for area code 415 and 510 the premium customers that churned are 16 and 11 these should be arrested first.

Churn	Area Code	Count of Id	Prem Customer
False	408	716	11
	415	1,419	32
	510	715	13
True	408	122	9
	415	236	16
	510	125	11

Inference 2: The Company should devise a plan a better plan for premium customers, they should either start with discounting customers who take both international and voice mail plan.

8. The Evening average charge for churn customers is higher than the average customer who does not churn. This can be seen in the table below.

Churn	Area Code	Eve Avg ..	Avg. Total Eve Charge	Count of Id	Total Eve Calls	Total Eve Charge	Total Eve Min
False	408	17.1093..	17	716	71,614	12,144	142,875
	415	17.0555..	17	1,419	142,455	23,964	281,921
	510	17.1129..	17	715	71,041	12,111	142,478
True	408	17.1093..	18	122	12,009	2,193	25,802
	415	17.0555..	18	236	23,879	4,263	50,159
	510	17.1129..	18	125	12,683	2,264	26,634

Similarly the day charge for a churn customer is higher than the average customer.

Churn	Area Code	Day Avg Charg..	Avg. Total Day Char..	Count of Id	Total Day Calls	Total Day Charge	Total Day Minutes
False	408	30.120274463	29	716	71,673	20,949	123,229
	415	30.871335347	30	1,419	142,761	42,648	250,867
	510	30.394428571	30	715	71,373	21,277	125,155
True	408	30.120274463	35	122	12,543	4,292	25,244
	415	30.871335347	36	236	23,693	8,444	49,669
	510	30.394428571	34	125	12,709	4,255	25,027

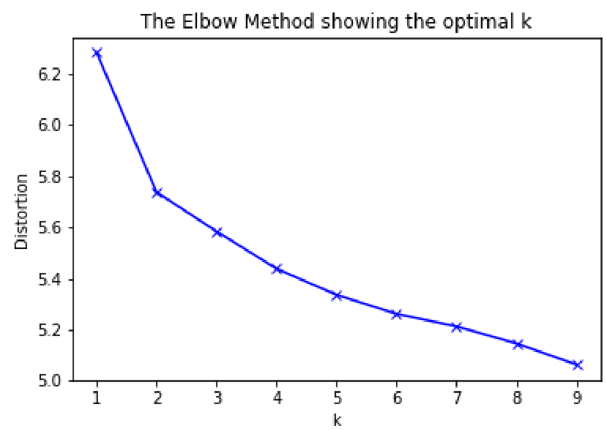
Churn	Area Code	Day Avg Charg..	Avg. Total Day Char..	Count of Id	Avg. Total Day Calls	Avg. Total Day Min..
False	408	30.120274463	29	716	100	172
	415	30.871335347	30	1,419	101	177
	510	30.394428571	30	715	100	175
True	408	30.120274463	35	122	103	207
	415	30.871335347	36	236	100	210
	510	30.394428571	34	125	102	200

Inference 3: The day charge is significantly higher, company should identify high value customers who make significantly higher calls than a threshold and should have a new call time plan for day

Stage: 3 Mental Model Preparation

This is the technique where validation of customer behavior was required, developed an unsupervised clustering algorithm to validate if the model too is able to verify two classes in the data set. In the adjacent graph we can see that the curve starts to bend at k=2, this shows that there is some pattern in the data that defines two or three clusters which is true according to our hypothesis about our problem.

Inference 4: This builds into our hypothesis, that the data represents two to three classes.



Stage 5: Validation Metric

For this model evaluation we are going to use the below mentioned validation metrics:

a. Accuracy Score
b. Recall
c. F1 Score
d. Confusion Matrix
e. ROC-AUC Curve

		Predicted class	
Actual Class		Class = Yes	Class = No
	Class = Yes	True Positive	False Negative
	Class = No	False Positive	True Negative

True Positives (TP) - These are the correctly predicted positive values which means that the value of actual class is yes and the value of predicted class is also yes.
True Negatives (TN) - These are the correctly predicted negative values which means that the value of actual class is no and value of predicted class is also no.
False positives and false negatives, these values occur when your actual class contradicts with the predicted class.
False Positives (FP) – When actual class is no and predicted class is yes.
False Negatives (FN) – When actual class is yes but predicted class in no.

Accuracy - Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. One may think that, if we have high accuracy then our model is best. Yes, accuracy is a great measure but only when you have symmetric datasets where values of false positive and false negatives are almost same. Therefore, you have to look at other parameters to evaluate the performance of your model. For our model, we have got 0.803 which means our model is approx. 80% accurate.
Accuracy = TP+TN/TP+FP+FN+TN

Precision - Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. High precision relates to the low false positive rate.
Precision = TP/TP+FP

Recall (Sensitivity) - Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes. The question recall answers is
Recall = TP/TP+FN

F1 score - F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it's better to look at both Precision and Recall.
F1 Score = 2*(Recall * Precision) / (Recall + Precision)

Stage 6: Features
Based on business intelligence, developed set of features – Appendix

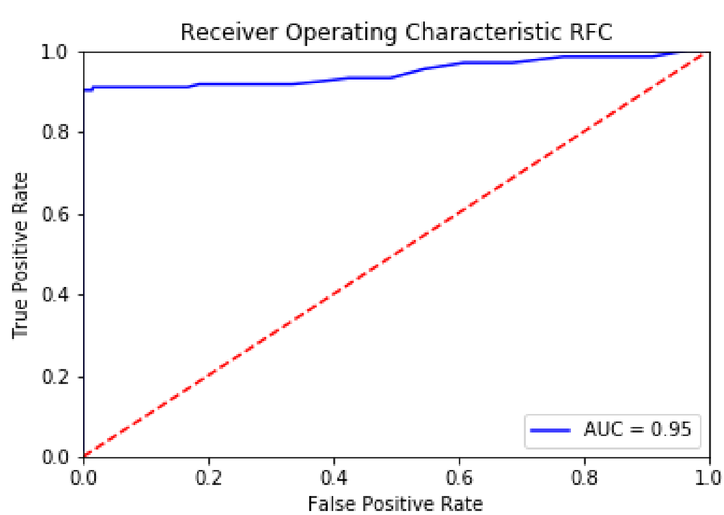
Stage 7: ML Models and Results || Selection – Validation

Classifiers	Recall	TP	TN	FP	FN	F1	Accuracy	AUC
KNeighborsClassifier	0.34	45	852	15	88	0.47	0.89	80
RandomForestClassifier	0.86	114	867	0	19	0.92	0.98	95
GaussianNB	0.66	89	763	104	44	0.54	0.85	83
LogisticRegression	0.24	32	832	35	101	0.32	0.86	81
LogisticRegressionCV	0.22	30	834	33	103	0.3	0.86	81
DecisionTreeClassifier with GINI	0.89	119	844	23	14	0.85	0.96	93
DecisionTreeClassifier with Enropy	0.9	120	830	37	12	0.85	0.95	93
SVC	0.44	59	857	10	74	0.58	0.91	91
Gradient Boosting Classifier	0.39	52	848	19	81	0.51	0.9	0.51

Used Grid Search to tune the hyper parameters with classifier as the Radom Forest, the result is mentioned below:

Classifier	Recall	TP	TN	FP	FN	F1	Accuracy	AUC
Random Forest with Grid Search	0.72	96	864	3	37	0.82	0.96	94

Best Model – RandomForestClassifier: RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini', max_depth=100, max_features='auto', max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, n_estimators=120, n_jobs=20, oob_score=False, random_state=123, verbose=0, warm_start=False)



7. Data Informed business solution

   Strategy for Business:

   1. Used the model developed above to identify the customers which can be potential churners.
   2. Identify those customers which use premium plans and build discount plans for using all the premium plans.
   3. The operator should identify the call minutes of potential churners with the rest and should revise their minute charge after a particular threshold. People who make calls are revenue generators for the company hence the company should not lose them. We have seen above that there the people who have churned have higher call charge because their call minutes are more, so decision makers can also decide to launch a plan which would help those people who have a certain threshold of call minutes

8. Appendix

1. Feature Engineered

Feature	Meaning
Premium Customer	Customer who avails both voice mail plan and international plan
day_avg_charge	average charge for day calls
eve_avg_charge	average charge for evening calls
night_avg_charge	average charge for night calls
total_calls	total customer call count
total_charge	total customer charge
total_min	total customer minutes
day_avg_charge[data]	avg day charge average
eve_avg_charge[data]	avg eve charge average
night_avg_charge[data]	avg night charge average
day_avg_charge[data_state]	avg day charge in the state
eve_avg_charge[data_state]	avg eve charge in the state
night_avg_charge[data_state]	avg night charge in the state
day_avg_charge[data_area_code]	average day charge in the area code
eve_avg_charge[data_area_code]	average eve charge in the area code
night_avg_charge[data_area_code]	average night charge in the area code
deviation_day_charge_state	deviations
deviation_eve_charge_state	deviations
deviation_night_charge_state	deviations
deviation_day_charge_area_code	deviations
deviation_eve_charge_area_code	deviations
deviation_eve_charge_area_code	deviations
cus_dev_day_charge	deviations
cus_dev_eve_charge	deviations
cus_dev_night_charge	deviations

2. Correlation and Distribution Plot