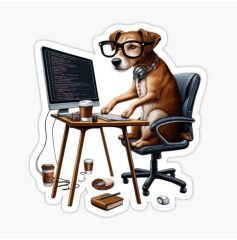
Driving Business Impact: Reducing Churn and Enhancing Market Strategy with IBM's Telco Dataset

IBM Telco Data: Customer Churn Analysis

Yogesh Gupta
Abhiram
John
Dozie

Problem Statement and Objectives



Customer churn is a major challenge for telecommunications companies, resulting in revenue loss, increased operational costs, and, in some extreme cases, business closures—such as the case of Aircel. This project aims to identify the key factors contributing to churn and develop retention strategies to ensure the company remains competitive. Additionally, the project focuses on optimizing marketing strategies and refining customer segmentation, aiming to improve customer satisfaction, increase loyalty, and ensure sustainable growth for the company.

The primary objectives of this project are as follows:

1. Predict Churn: Identify customers most likely to churn in order to take proactive measures to save operational costs related to customer acquisition.
2. Understand Drivers of Churn: Gain insights into the key factors contributing to churn, allowing for the development of effective retention strategies.
3. Target Marketing Optimization: Categorize customers based on churn risk to enable targeted marketing and personalized offers.
4. Provide Business Insights: Provide actionable insights that will help the company maintain profitability and improve decision-making.

Key performance indicators on our radar are : 1. reducing the churn rate by offering the recommended retention strategies 2. Increasing the customer tenure. 3. Reducing the new customer acquisition costs, as retaining existing customers is cheaper than acquiring new ones.

Data Preparation and Exploration

1. **Dataset Overview** - 1. The dataset used was from Kaggle through it's APIs provided., contains 7043 rows and 21 columns, with customer churn as the target variable.
2. **Data Cleaning Process**:

- A) **Data Cleaning Process**
- i) Dropped duplicates.
- ii)Addressed empty rows in the 'TotalCharges' column- there were empty rows
- iii) Corrected data types and handled categorical variables.
- iv) Inputed missing values using median due to skewness (0.96)- median -1397v)
- v) Saved the cleaned data as "cleaned_stage1.csv".
- i) Searched for outliers through box-plots but no outliers were found in "MonthlyCharges", :"tenure", "TotalCharge

```
Column 'customerID' - Blank values: 0
Column 'gender' - Blank values: 0
Column 'SeniorCitizen' - Blank values: 0
Column 'Partner' - Blank values: 0
Column 'Dependents' - Blank values: 0
Column 'tenure' - Blank values: 0
Column 'PhoneService' - Blank values: 0
Column 'MultipleLines' - Blank values: 0
Column 'InternetService' - Blank values: 0
Column 'OnlineSecurity' - Blank values: 0
Column 'OnlineBackup' - Blank values: 0
Column 'DeviceProtection' - Blank values: 0
Column 'TechSupport' - Blank values: 0
Column 'StreamingTV' - Blank values: 0
Column 'StreamingMovies' - Blank values: 0
Column 'Contract' - Blank values: 0
Column 'PaperlessBilling' - Blank values: 0
Column 'PaymentMethod' - Blank values: 0
Column 'MonthlyCharges' - Blank values: 0
Column 'TotalCharges' - Blank values: 11
Column 'Churn' - Blank values: 0
```

3. **Exploratory Data Analysis (EDA)**:

Conducted visualizations including line charts and histograms for tenure and churn.

Analyzed churn rate across various customer segments (e.g., senior citizens, internet service type).

Identified key features influencing churn, such as contract type, payment method, and internet service type.

Used correlation analysis to identify relationships between features.

Category	Total Customers	Churned Customers	Churn Percentage
Male	3555	930	26.16%
Female	3488	939	26.92%
Non-Senior	5901	1393	23.61%
Senior	1142	476	41.68%

Contract	Month-to-month	1 year	2 years
Total customers	3875	1473	1695
churned	1655	166	48
Churned %	42.70967	11.269518	2.831858

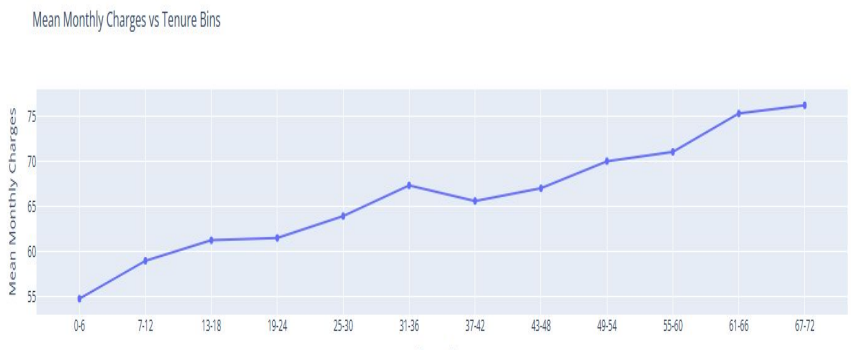
Payment type	Bank transfer (automatic)	Credit card (automatic)	Electronic check	Mailed check
total_customers	1544	1522	2365	1612
churned_customers	258	232	1071	308
churn_rate	16.7%	15.24	45.28%	19.10%

Exploratory Data Analysis

- Visualized churn rates across different tenure intervals (e.g., 0-3 months, 3-6 months, etc.).
- Grouped and calculated data for **3-month intervals** showing mean, median, minimum, and maximum **monthly charges** paid by customers.
- Found that the **churn rate** was highest among early customers (0-12months with the company).
- Identified correlations between various columns (e.g., **Monthly Charges**, **Tenure**) and churn.
- Detected anomalies with **0 tenure** but non-zero **TotalCharges**.
- **Senior citizens** (both males and females) who **do not use Tech Support or Online Security** exhibit **higher churn** rates.
- Investigated and found it was due to a **system issue** where the tenure was not updated in the database.Customers with NaN **TotalCharges** were imputed with the median value. All these customers had **Churn = 0**, indicating they are **valid, active customers**.
- **Generated Visualizations**

	tenure_bins_3_months	mean_monthly_charges	median_monthly_charges	min_monthly_charges	max_monthly_charges	customer_count	churn_percentage
0	0-2	52.225638	50.450	18.75	104.40	862	58.352668
1	3-5	58.594597	64.350	18.80	107.95	509	47.347741
2	6-8	57.909615	61.150	18.95	109.90	364	36.538462
3	9-11	60.052844	66.075	18.85	111.40	334	36.526946
4	12-14	59.470695	61.725	18.80	112.95	302	33.112583
5	15-17	61.854511	69.300	18.80	112.95	266	34.210526
6	18-20	59.456224	61.500	18.80	108.20	241	25.311203
7	21-23	62.903361	68.700	19.05	111.20	238	23.949580
8	24-26	61.848810	66.175	18.70	108.90	252	24.206349
9	27-29	63.062687	68.850	18.25	110.85	201	19.900498
10	30-32	70.688350	79.375	18.95	110.45	206	24.757282
11	33-35	65.570046	70.300	19.15	116.25	217	18.894009
12	36-38	65.313793	73.150	18.55	110.70	174	21.839080
13	39-41	65.852368	70.225	19.10	114.50	190	21.578947
14	42-44	67.132597	73.850	19.05	115.05	181	19.337017
15	45-47	67.303202	74.800	18.85	115.65	203	15.763547
16	48-50	69.679545	75.350	19.00	117.45	198	17.171717
17	51-53	67.492202	75.900	18.70	111.80	218	13.761468
18	54-56	72.347406	78.600	18.95	116.50	212	15.094340
19	57-59	69.454687	74.475	18.40	113.75	192	14.062500
20	60-62	73.182432	81.025	19.10	118.60	222	8.558559
21	63-65	76.741886	85.375	19.15	115.10	228	7.456140
22	66-68	73.834495	80.450	19.25	118.60	287	11.149826
23	69-72	77.163874	85.450	19.10	118.75	746	4.155496

```
Correlation between MonthlyCharges and Tenure:
               MonthlyCharges  tenure
MonthlyCharges          1.0000  0.2479
tenure                  0.2479  1.0000
```



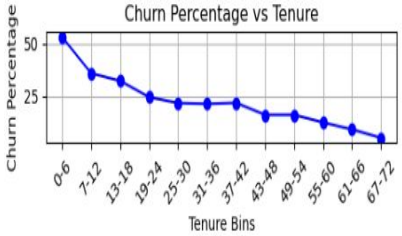
	tenure_bins	count	mean_monthly_charges	median_monthly_charges	min_monthly_charges	max_monthly_charges	customer_count	churn_percentage
0	0-6	1481	54.738656	54.700	18.75	109.90	1481	52.937205
1	7-12	705	58.952908	64.950	18.85	112.95	705	35.886525
2	13-18	548	61.236131	68.475	18.80	112.95	548	32.299270
3	19-24	476	61.496744	65.150	18.80	111.20	476	24.579832
4	25-30	431	63.929698	69.750	18.25	110.85	431	21.809745
5	31-36	401	67.344389	75.200	18.55	116.25	401	21.446384
6	37-42	379	65.610158	70.600	19.05	114.50	379	21.899736
7	43-48	383	67.018930	73.850	18.85	117.45	383	16.187990
8	49-54	420	70.039286	79.175	18.70	115.60	420	16.190476
9	55-60	412	71.072209	76.800	18.40	116.60	412	12.621359
10	61-66	463	75.335529	84.500	19.10	118.60	463	9.287257
11	67-72	944	76.255403	84.800	19.10	118.75	944	5.296610

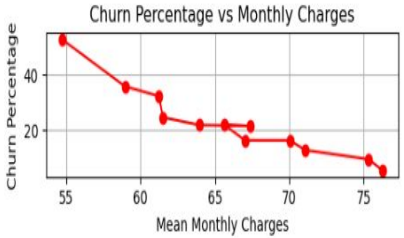
EDA Continued

- Male senior citizens were having more difficulty (52%) in mailing checks than females of same age criteria (42%).
- Senior citizens have almost 3X more churning rate when they use mailed checks than non senior citizens.
- Electronic check users, especially among both senior and non-senior citizens, have a significantly higher churn rate compared to other payment methods, with some segments reaching over 50% churn (e.g., senior citizens).
- Senior citizens on Month-to-month contracts churned at 53%, while non-senior citizens churned at 39% for the same contract type.
- senior citizens on One year and Two year contracts show lower churn rates compared to non-senior citizens in these categories.
- Senior females have a higher churn rate for StreamingTV (47.83%) compared to senior males (39.11%).

Customer Segments to target?

- Early customers
- Senior citizens- both male and females, offering specialized services for those who use fiber optic services.
- Give incentives to customers to switch from month-to month to yearly contracts.
- Regular customer feedback, loyalty programs missing
- Make payment processes smoother, check with the team handling checks.
- Offer free trials to customers to use online security and tech support. Also, can make the system user friendly.
- Offer bundled offers to customers to use streamingTV and streamingMovies so as to make them churn less.





Supervised Machine Learning Models

- Model Selection - Used 5 machine learning models to test and train the data and check which model is best. Used Supervised learning as the data has binary data columns and target column- churn which is best suited for supervised learning. I used Logistic regression, Decision tree classifier, Random Forest, gradient boosting, and a hybrid model (LR, DC,RF,KNN, SVM) to test and model the dataset. The model selection process involved comparing precision and recall across models, balancing the need to capture churners while minimizing false positives. For example, Gradient Boosting prioritized recall at the cost of precision, while Random Forest and Logistic Regression balanced both metrics better.
- Trained and tested the dataset in 80:20 ratio.
- trained the model with a default threshold of 0.5 for predicting churn.
- Evaluated performance using accuracy, classification report, and confusion matrix.
- Used 5-fold cross-validation to assess the model's consistency and calculate the mean cross-validation accuracy.
- Identified key features influencing customer churn based on the model's coefficients.
- Hyperparameter Tuning
- Plotted the ROC curve and calculated the AUC to measure the model's ability to distinguish between churners and non-churners.
- Plotted the precision-recall curve, calculated the best threshold, and selected the optimal decision boundary.
- Applied the best threshold to adjust the decision-making process and improved model performance.
- After retraining with the custom threshold, we re-evaluated performance with accuracy and cross-validation, ensuring the model's robustness.

Key Highlights :

A) ACCURACY

- Decision Tree and Random Forest models show very high training accuracy (99.82%), indicating overfitting in their default forms.
- Gradient Boosting has lower training accuracy (83.05%), but it is more reasonable compared to Decision Tree and Random Forest.
- Logistic Regression has a similar training accuracy to Gradient Boosting but is much lower than the decision tree and random forest models.

B) Comparison

- Gradient Boosting (Default): 78.92% | Tuned Best Threshold: 74.95%
- Random Forest (Default): 78.07% | Tuned Best Threshold: 79.15%
- Decision Tree (Default): 72.46% | Pruned: 77.71%
- Logistic Regression (Default): 79.77% | Tuned Best Threshold: 79.63%
- Gradient Boosting: High recall (81.09%) at the cost of low precision (52.78%) after tuning.
- Random Forest: Improved precision (61.42%) after tuning with moderate recall improvement.
- Decision Tree: Precision increases (63.95%) after pruning, but recall decreases (42.75%).

Metric	Logistic Regression (Tuned E	Decision Tree (Pruned Best Threshold)	Random Forest (Tuned Best Threshold)	Gradient Boosting (Tuned Best Threshold)
Training Accuracy	80.86%	99.82%	99.82%	83.05%
Testing Accuracy	79.63%	77.71%	79.15%	74.95%
Precision (Class 1)	54.59%	51.15%	61.42%	52.78%
Recall (Class 1)	76.94%	74.87%	72.94%	81.09%
F1-Score (Class 1)	63.87%	60.78%	66.26%	63.94%
AUC (ROC)	0.85	0.84	0.88	0.87
AUC (Precision-Recall)	0.8	0.79	0.84	0.82
Best Threshold	0.28	0.4	0.35	0.32

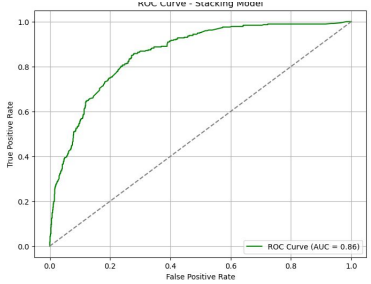
Model	Best For	Use Case in Telco R	When to Use
Logistic Regression (Tuned Best Threshold)	Balanced accuracy and simplicity	Good for general predictions when interpretability is needed, and a balance between	When you need a model that performs adequately without being
Decision Tree (Pruned Best Threshold)	Interpretability and quick insights	Quick insights and simple models that provide clear decision-making paths.	When you need a model that gives clear decision rules or insights into customer behavior
Random Forest (Tuned Best Threshold)	High precision and balanced performance	Ideal for customer retention where precision (minimizing false positives) is key	When reducing false negatives (churners) is important, and you need a balanced
Gradient Boosting (Tuned Best Threshold)	Maximizing recall (catching more churners)	Best for predicting churn, especially when the focus is to catch as many churners as	When you are most concerned with catching churners (high recall), even if it means dealing

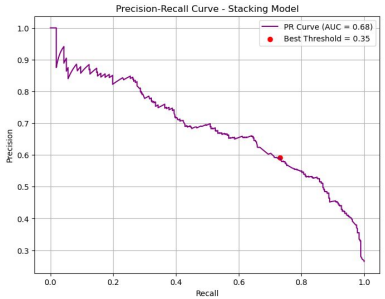
Hybrid model- accuracy of training and testing= 82.53%, 80.90%

- Precision for class 0 is high (0.85), but for class 1, recall (0.73) is prioritized over precision (0.59), making it good for identifying churn but with some trade-offs in false positives.
- The model correctly identifies a majority of churners (true positives) and non-churners (true negatives), with fewer false positives and false negatives.
- True Negatives (932), False Positives (104),False Negatives (165),True Positives (208)
- The stacking model shows improved performance over individual base models, with **strong area under the curve AUC (0.86) and recall**
- **PR curve on hovering will show how on different thresholds precision and recall vary. Best is at 0.35**

```
Best Threshold for Stacking Model: 0.3502059799381969
Training Accuracy with Custom Threshold: 0.825346112886049
Testing Accuracy with Custom Threshold: 0.794889992902768
```

```
Confusion Matrix (Stacking Model):
[[932 104]
 [165 208]]
```



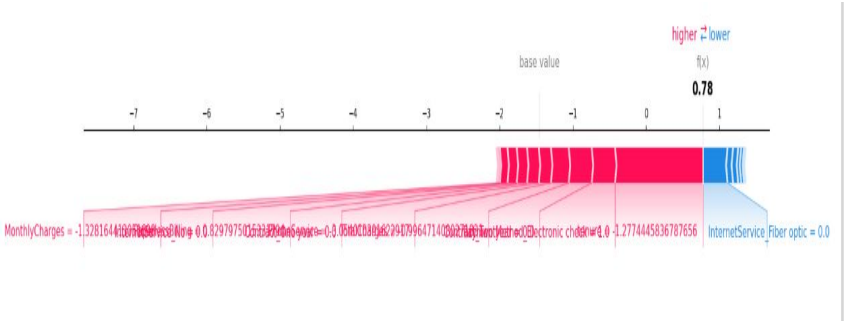


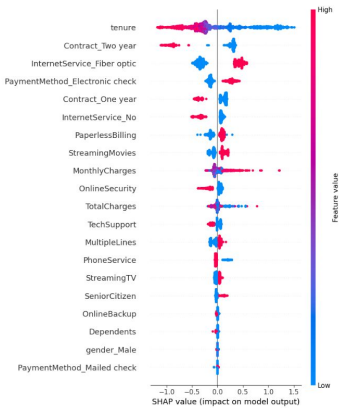
Features, Interpretability techniques - **Shap & Lime**

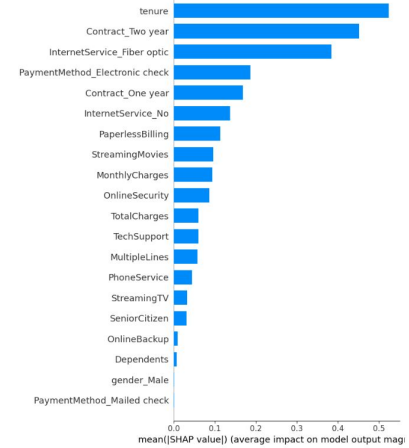
Main features across all models that affect churn:

1. Tenure - Appears in all models (Logistic Regression, Decision Tree, Random Forest, Gradient Boosting). It consistently shows a strong relationship with churn, with longer tenure generally reducing the likelihood of churn.
2. InternetService_Fiber optic-This feature also appears in all models as an important predictor of churn, with customers on Fiber optic services more likely to churn.
3. PaymentMethod_Electronic check: This feature is an important predictor across Logistic Regression, Random Forest, Gradient Boosting, and Tuned Gradient Boosting, with electronic check payment method correlating with higher churn likelihood.
4. Contract Types: Two-year contracts are protective against churn, while one-year contracts increase churn likelihood.
5. TotalCharges & MonthlyCharges: This feature appears in multiple models and plays a critical role in churn prediction, with higher total charges often associated with a higher likelihood of churn.

SHAP analysis on tuned gradient boosting. The red color shows parameter reducing churn while blue shows increasing churn rate. tenure, Contract_Two year, and InternetService_Fiber optic show a significant impact, both in positive and negative directions, as reflected in their SHAP values.





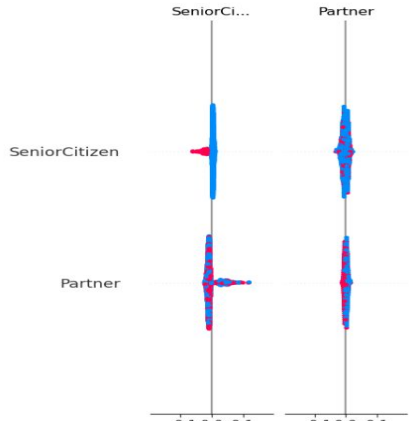


Higher MonthlyCharges may not always lead to churn, as seen in the negative SHAP value, but InternetService_Fiber optic on the positive side indicates a need for improved service. To reduce churn, the company should focus on enhancing customer support and aligning fiber optic service quality with

Features, Interpretability techniques - **Shap & Lime continued**

- Random forest model used - distribution of SeniorCitizen is tight, showing a relatively small but consistent effect on the churn prediction. The blue points (indicating lower values for SeniorCitizen) suggest that non-senior customers have a higher likelihood of churn, whereas senior customers (represented by the red dots) are less likely to churn, with a clear negative SHAP value. This analysis is contradictory to the EDA which might mean Seniors might churn early on (in the first 0-6 months), leading to a higher initial churn rate.
- Senior citizens are more likely to stay as they show a **negative contribution to churn**. The SHAP values for seniors are **mostly positive**
- **No Partner** (indicated by **blue dots**) appears to have a more **negative SHAP value**, pushing churn predictions **higher**
- Customers with a **Partner** (indicated by **red dots**) tend to have **positive SHAP values**, meaning having a partner **lowers the churn likelihood**.

1. PaperlessBilling and Contract_Two year appear to increase churn, meaning that customers with paperless billing and shorter contracts are more likely to churn. (REASONS???) - less connected, tech savvy- better deals, difficulty in keeping track of payments
2. MonthlyCharges, InternetService_No, and PaymentMethod_Electronic check work to decrease churn for this particular customer.
3. Electronic check is in blue, which means that not using EC lowers the churn probability. And contract_two_year=0 meaning customer not using this and by not having a two-year contract is increasing the likelihood of churn.



Recommendations for the company

1. **Improve Customer Service for Fiber Optic Customers**: Both SHAP and LIME indicate that Fiber optic internet correlates with higher churn, likely due to customer dissatisfaction or poor service quality.

   Solution : Invest in improving the fiber optic service with better customer support, faster issue resolution, and service upgrades. Provide dedicated customer service for fiber optic users, offer troubleshooting help, and ensure service reliability.

   Estimated impact : Reduced churn in fiber optic users, increasing retention and overall satisfaction.

2. **Incentivize Customers to Choose Two-Year Contracts-** EDA on dataset, features extracted PDPs , SHAP and LIME, all show that two-year contracts correlate with lower churn, while one-year contracts and no contract tend to increase churn likelihood.

   Solution : - Have discounts on contracts to influence customers to move from monthly subscriptions. Create campaigns that highlight the benefits of longer-term contracts, emphasizing cost savings and added benefits.

   Expected Impact: Increased contract retention and reduced churn, locking in customers for a longer period.

3. Enhance Payment Method Flexibility: Electronic checks have shown to be one of the main features in our models and SHAP analysis for high churn rates.

   Solution - Tie up with banks to offer cashback and discounts on the bank's credit/debit cards. Offer incentives (e.g., small discounts or loyalty points) to encourage the use of automatic payment methods. Set up a payment monitoring system to reduce delayed payments, and promote automated payments.

   Expected Impact: Increased payment retention and decreased churn, particularly from customers using less reliable payment methods.

4. **Prioritize Early-Stage Customer Retention (0-12 Months):** SHAP , EDA suggest that customers in the first year have a higher likelihood of churn, especially if they are dissatisfied early on. So get them on contract as soon as possible. Focus on early engagement for customers in the 0-12 months category to convert them into long-term customers. Onboarding programs to engage new customers right after sign-up, including tutorials, personalized communication, and onboarding offers. Take regular customer feedback and analyze the common issues they are facing.

   Expected Impact: Higher early-stage retention, leading to more long-term customers and a reduced churn rate during the critical early phase.

5. Create loyalty programs or special benefits for customers who have been with the company for longer periods (e.g., 1-2 years). This can increase retention for customers already showing loyalty.
6. Engage customers in using more and more services. This can be done by giving free trials or discounted offers to creat a sense of habit forming in new customers.
7. Price Sensitivity: Use segmentation to tailor pricing strategies based on customer willingness to pay. This is due to SHAP showing counter-intuitive results,