Telco Customer Project

Andrew Lee, Hisham Galib

University of Texas at Austin

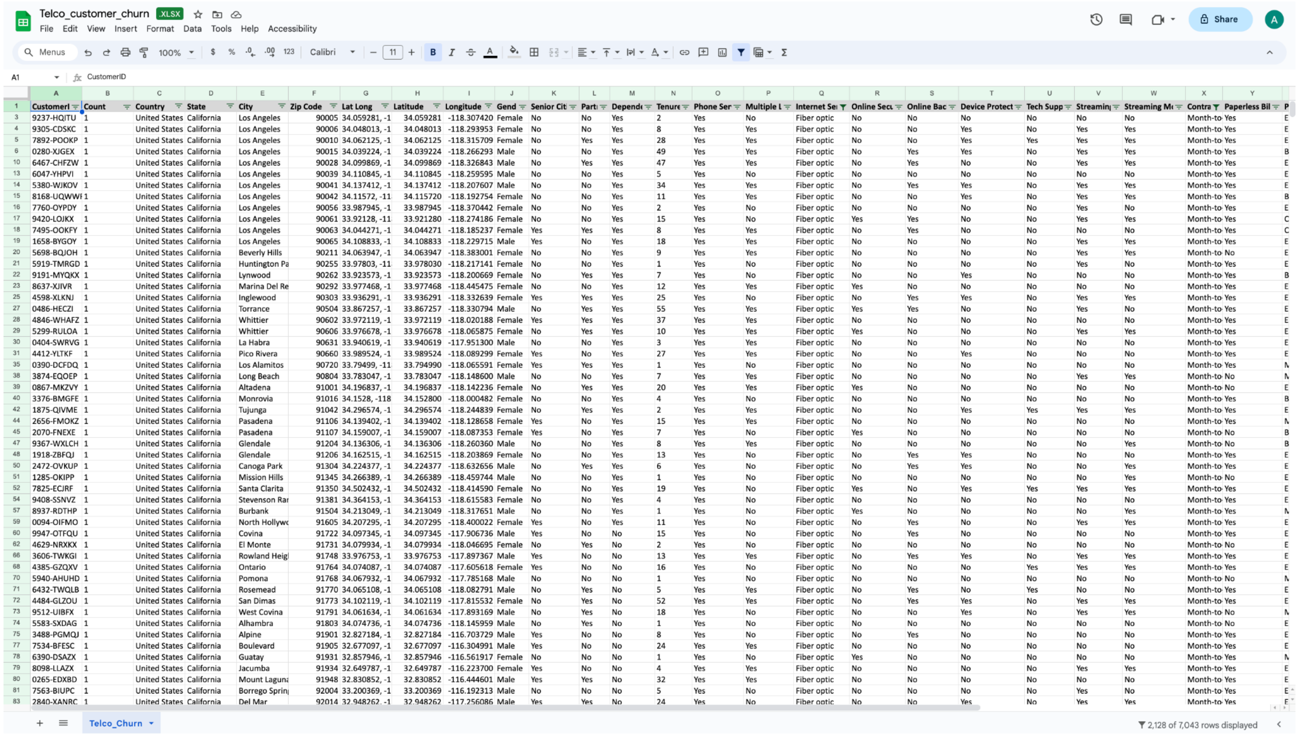
Table of Contents

Introduction

With growing competition from big companies such as Verizon and AT&T, telecom providers constantly compete with one another to attract more customers, often taking customers from the other provider with incentives promised to the customer after switching providers. The situation described is related to what is referred to as 'churn' in the telecommunication industry. Churn is the measure of how many customers stop using a product or service. Customer retention has been an imperative focus from telecom providers to sustain revenue, attempt to foster lasting relationships, and ensure customer satisfaction. With competition growing constantly and providers finding new ways to attract customers as well as retain customers, one of the best ways to go about finding methods to accomplish mitigating churn and potentially finding new customers is to analyze patterns of churn, uncover insights, and proposing strategies based off those patterns and insights. Demonstrating these techniques in the Telco Customer Churn dataset, a synthetic dataset produced by IBM to model a hypothetical telecommunications company. With the dataset, analyzing the data to find out if the customer is at risk of churn or is undergoing churn, and finding out how to retain customers through a proposed solution based off the analysis is the goal. Using key features such as 'Payment Method', 'Monthly Charges', and 'Internet Service' as the leading factors in increasing churn, techniques such as an Exploratory Data Analysis (EDA) and Synthetic Minority Oversampling Technique (SMOTE) were used to decide how much these features affected churn and how they can be harnessed to reduce churn as well as to retain customers.

Description of Data

The Telco Customer Churn dataset is a repository of information, providing data on the behaviors and preferences of customers. The dataset includes key attributes that define customers, ranging from demographic details to the type of service they subscribe to including the main variable ‘Churn’ which indicates whether the customer has chosen to discontinue or continue their services with Telco. The dataset can be found here.

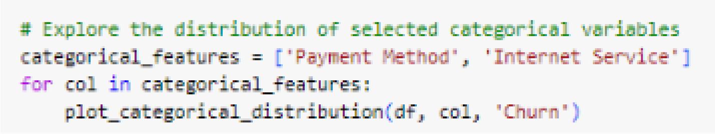


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	
1	Customerl	Count	Country	State	City	Zip Code	Lat Long	Latitude	Longitude	Gend	Senior Citi	Parti	Depende	Tenure	Phone Ser	Multiple L	Internet Se	Online Secu	Online Bac	Device Protect	Tech Supp	Streamin	Streaming M	Contra	Paperless Bil	P
3	9237-HQITU	1	United States	California	Los Angeles	90005	34.059281, -1	34.059281	-118.307420	Female	No	No	Yes	2	Yes	No	Fiber optic	No	No	No	No	No	No	Month-to	Yes	E
4	9305-CDSKC	1	United States	California	Los Angeles	90006	34.048013, -1	34.048013	-118.293953	Female	No	No	Yes	8	Yes	Yes	Fiber optic	No	No	Yes	No	Yes	Yes	Month-to	Yes	E
5	7892-POOKP	1	United States	California	Los Angeles	90010	34.062125, -1	34.062125	-118.315709	Female	No	Yes	Yes	28	Yes	Yes	Fiber optic	No	No	Yes	Yes	Yes	Yes	Month-to	Yes	E
6	0280-XJGEX	1	United States	California	Los Angeles	90015	34.039224, -1	34.039224	-118.266293	Male	No	No	Yes	49	Yes	Yes	Fiber optic	No	Yes	Yes	No	Yes	Yes	Month-to	Yes	B
10	6467-CHFZW	1	United States	California	Los Angeles	90028	34.099869, -1	34.099869	-118.326843	Male	No	Yes	Yes	47	Yes	Yes	Fiber optic	No	Yes	No	No	Yes	Yes	Month-to	Yes	E
13	6047-YHPVI	1	United States	California	Los Angeles	90039	34.110845, -1	34.110845	-118.259595	Male	No	No	Yes	5	Yes	No	Fiber optic	No	No	No	No	No	No	Month-to	Yes	E
14	5380-WJKOV	1	United States	California	Los Angeles	90041	34.137412, -1	34.137412	-118.207607	Male	No	No	Yes	34	Yes	Yes	Fiber optic	No	Yes	Yes	No	Yes	Yes	Month-to	Yes	E
15	8168-UQWWI	1	United States	California	Los Angeles	90042	34.11572, -11	34.115720	-118.192754	Female	No	No	Yes	11	Yes	Yes	Fiber optic	No	No	Yes	No	Yes	Yes	Month-to	Yes	B
16	7760-OYPDY	1	United States	California	Los Angeles	90056	33.987945, -1	33.987945	-118.370442	Female	No	No	Yes	2	Yes	No	Fiber optic	No	No	No	No	Yes	No	Month-to	Yes	E
17	9420-LOJKX	1	United States	California	Los Angeles	90061	33.92128, -11	33.921280	-118.274186	Female	No	No	Yes	15	Yes	No	Fiber optic	Yes	Yes	No	No	Yes	Yes	Month-to	Yes	C
18	7495-OOKFY	1	United States	California	Los Angeles	90063	34.044271, -1	34.044271	-118.185237	Female	Yes	Yes	Yes	8	Yes	Yes	Fiber optic	No	Yes	No	No	No	No	Month-to	Yes	C
19	1658-BYGOY	1	United States	California	Los Angeles	90065	34.108833, -1	34.108833	-118.229715	Male	Yes	No	Yes	18	Yes	Yes	Fiber optic	No	No	No	No	Yes	Yes	Month-to	Yes	E
20	5698-BQJOH	1	United States	California	Beverly Hills	90211	34.063947, -1	34.063947	-118.383001	Female	No	No	Yes	9	Yes	Yes	Fiber optic	No	No	No	No	Yes	Yes	Month-to	No	E
21	5919-TMRGD	1	United States	California	Huntington Pa	90255	33.97803, -11	33.978030	-118.217141	Female	No	No	Yes	1	Yes	No	Fiber optic	No	No	No	No	Yes	No	Month-to	Yes	E
22	9191-MYQKX	1	United States	California	Lynwood	90262	33.923573, -1	33.923573	-118.200669	Female	No	Yes	Yes	7	Yes	No	Fiber optic	No	No	Yes	No	No	No	Month-to	Yes	B
23	8637-XJIVR	1	United States	California	Marina Del Re	90292	33.977468, -1	33.977468	-118.445475	Female	No	No	Yes	12	Yes	Yes	Fiber optic	Yes	No	No	No	No	No	Month-to	Yes	E
25	4598-XLKNJ	1	United States	California	Inglewood	90303	33.936291, -1	33.936291	-118.332639	Female	Yes	Yes	Yes	25	Yes	No	Fiber optic	No	Yes	Yes	No	Yes	Yes	Month-to	Yes	E
27	0486-HECZI	1	United States	California	Torrance	90504	33.867257, -1	33.867257	-118.330794	Male	No	Yes	Yes	55	Yes	Yes	Fiber optic	Yes	Yes	No	No	Yes	No	Month-to	Yes	E
28	4846-WHAFZ	1	United States	California	Whittier	90602	33.972119, -1	33.972119	-118.020188	Female	Yes	Yes	Yes	37	Yes	Yes	Fiber optic	No	No	No	No	No	No	Month-to	Yes	E
29	5299-RULOA	1	United States	California	Whittier	90606	33.976678, -1	33.976678	-118.065875	Female	No	Yes	Yes	10	Yes	Yes	Fiber optic	Yes	No	No	No	Yes	Yes	Month-to	Yes	E
30	0404-SWRVG	1	United States	California	La Habra	90631	33.940619, -1	33.940619	-117.951300	Male	No	No	Yes	3	Yes	Yes	Fiber optic	No	No	No	No	No	No	Month-to	Yes	E
31	4412-YLTKF	1	United States	California	Pico Rivera	90660	33.989524, -1	33.989524	-118.089299	Female	Yes	No	Yes	27	Yes	Yes	Fiber optic	No	No	Yes	No	No	No	Month-to	Yes	E
35	0390-DCFDQ	1	United States	California	Los Alamitos	90720	33.79499, -11	33.794990	-118.065591	Female	Yes	Yes	Yes	1	Yes	No	Fiber optic	No	No	No	No	No	No	Month-to	Yes	M
38	3874-EQOEP	1	United States	California	Long Beach	90804	33.783047, -1	33.783047	-118.148600	Male	No	No	Yes	7	Yes	Yes	Fiber optic	No	No	No	No	Yes	Yes	Month-to	No	M
39	0867-MKZVY	1	United States	California	Altadena	91001	34.196837, -1	34.196837	-118.142236	Female	No	Yes	Yes	20	Yes	Yes	Fiber optic	Yes	No	No	No	No	No	Month-to	No	E
40	3376-BMGFE	1	United States	California	Monrovia	91016	34.1528, -118	34.152800	-118.000482	Female	No	No	Yes	4	Yes	No	Fiber optic	No	No	No	No	No	No	Month-to	Yes	B
42	1875-QIVME	1	United States	California	Tujunga	91042	34.296574, -1	34.296574	-118.244839	Female	No	Yes	Yes	2	Yes	Yes	Fiber optic	No	No	Yes	Yes	Yes	Yes	Month-to	Yes	E
44	2656-FMOKZ	1	United States	California	Pasadena	91106	34.139402, -1	34.139402	-118.128658	Female	Yes	No	Yes	15	Yes	Yes	Fiber optic	No	No	No	No	No	No	Month-to	Yes	M
45	2070-FNEXE	1	United States	California	Pasadena	91107	34.159007, -1	34.159007	-118.087353	Female	Yes	No	Yes	7	Yes	No	Fiber optic	Yes	No	No	No	No	No	Month-to	No	B
47	9367-WXLCH	1	United States	California	Glendale	91204	34.136306, -1	34.136306	-118.260360	Male	No	No	Yes	8	Yes	Yes	Fiber optic	No	No	No	No	No	Yes	Month-to	No	B
48	1918-ZBFQJ	1	United States	California	Glendale	91206	34.162515, -1	34.162515	-118.203869	Female	No	No	Yes	13	Yes	No	Fiber optic	No	Yes	Yes	No	No	No	Month-to	Yes	E
50	2472-OVKUP	1	United States	California	Canoga Park	91304	34.224377, -1	34.224377	-118.632656	Male	No	Yes	Yes	6	Yes	No	Fiber optic	No	Yes	Yes	No	No	Yes	Month-to	Yes	E
51	1285-OKIPP	1	United States	California	Mission Hills	91345	34.266389, -1	34.266389	-118.459744	Male	No	No	Yes	1	Yes	No	Fiber optic	No	No	No	No	No	Yes	Month-to	No	E
52	7825-ECJRF	1	United States	California	Santa Clarita	91350	34.502432, -1	34.502432	-118.414590	Female	No	No	Yes	19	Yes	No	Fiber optic	Yes	No	Yes	Yes	Yes	Yes	Month-to	Yes	E
54	9408-SSNVZ	1	United States	California	Stevenson Rar	91381	34.364153, -1	34.364153	-118.615583	Female	No	No	Yes	4	Yes	No	Fiber optic	No	No	No	No	No	No	Month-to	Yes	E
57	8937-RDTHP	1	United States	California	Burbank	91504	34.213049, -1	34.213049	-118.317651	Male	No	No	Yes	1	Yes	No	Fiber optic	Yes	No	No	No	No	Yes	Month-to	Yes	M
59	0094-OIFMO	1	United States	California	North Hollywc	91605	34.207295, -1	34.207295	-118.400022	Female	Yes	No	Yes	11	Yes	No	Fiber optic	No	Yes	No	No	Yes	Yes	Month-to	Yes	E
60	9947-OTFQU	1	United States	California	Covina	91722	34.097345, -1	34.097345	-117.906736	Male	Yes	No	No	15	Yes	No	Fiber optic	No	Yes	No	No	No	No	Month-to	Yes	E
62	4629-NRXKX	1	United States	California	El Monte	91731	34.079934, -1	34.079934	-118.046695	Female	No	Yes	No	2	Yes	No	Fiber optic	No	No	No	No	No	No	Month-to	No	E
66	3606-TWKGI	1	United States	California	Rowland Heigl	91748	33.976753, -1	33.976753	-117.897367	Male	Yes	No	No	13	Yes	Yes	Fiber optic	No	Yes	Yes	No	Yes	Yes	Month-to	Yes	E
68	4385-GZQXV	1	United States	California	Ontario	91764	34.074087, -1	34.074087	-117.605618	Female	Yes	No	No	16	Yes	No	Fiber optic	No	No	No	Yes	Yes	Yes	Month-to	Yes	E
70	5940-AHUHD	1	United States	California	Pomona	91768	34.067932, -1	34.067932	-117.785168	Male	No	No	No	1	Yes	No	Fiber optic	No	No	No	No	No	No	Month-to	No	M
71	6432-TWQLB	1	United States	California	Rosemead	91770	34.065108, -1	34.065108	-118.082791	Male	No	Yes	No	5	Yes	Yes	Fiber optic	No	Yes	No	Yes	No	No	Month-to	Yes	E
72	4484-GLZOU	1	United States	California	San Dimas	91773	34.102119, -1	34.102119	-117.815532	Female	No	Yes	No	52	Yes	Yes	Fiber optic	No	Yes	Yes	No	Yes	Yes	Month-to	Yes	E
73	9512-UIBFX	1	United States	California	West Covina	91791	34.061634, -1	34.061634	-117.893169	Male	No	Yes	No	18	Yes	No	Fiber optic	No	No	Yes	No	Yes	Yes	Month-to	No	M
74	5583-SXDAG	1	United States	California	Alhambra	91803	34.074736, -1	34.074736	-118.145959	Male	No	Yes	No	1	Yes	No	Fiber optic	No	No	No	No	No	No	Month-to	No	E
75	3488-PGMQJ	1	United States	California	Alpine	91901	32.827184, -1	32.827184	-116.703729	Male	Yes	No	No	8	Yes	No	Fiber optic	No	Yes	No	No	No	No	Month-to	Yes	E
77	7534-BFESC	1	United States	California	Boulevard	91905	32.677097, -1	32.677097	-116.304991	Male	Yes	No	No	24	Yes	Yes	Fiber optic	No	No	No	No	No	No	Month-to	Yes	E
78	6390-DSAZX	1	United States	California	Guatay	91931	32.857946, -1	32.857946	-116.561917	Female	No	No	No	1	Yes	No	Fiber optic	Yes	No	No	No	No	No	Month-to	Yes	M
79	8098-LLAZX	1	United States	California	Jacumba	91934	32.649787, -1	32.649787	-116.223700	Female	Yes	No	No	4	Yes	Yes	Fiber optic	No	No	No	No	Yes	Yes	Month-to	Yes	E
80	0265-EDXBD	1	United States	California	Mount Laguna	91948	32.830852, -1	32.830852	-116.444601	Male	Yes	Yes	No	32	Yes	Yes	Fiber optic	No	No	No	No	No	No	Month-to	Yes	E
81	7563-BIUPC	1	United States	California	Borrego Sprin	92004	33.200369, -1	33.200369	-116.192313	Male	No	No	No	5	Yes	No	Fiber optic	No	No	No	No	Yes	No	Month-to	Yes	E
83	2840-XANRC	1	United States	California	Del Mar	92014	32.948262, -1	32.948262	-117.256086	Male	Yes	Yes	No	24	Yes	No	Fiber optic	No	No	Yes	No	Yes	Yes	Month-to	Yes	E

2,128 of 7,043 rows displayed

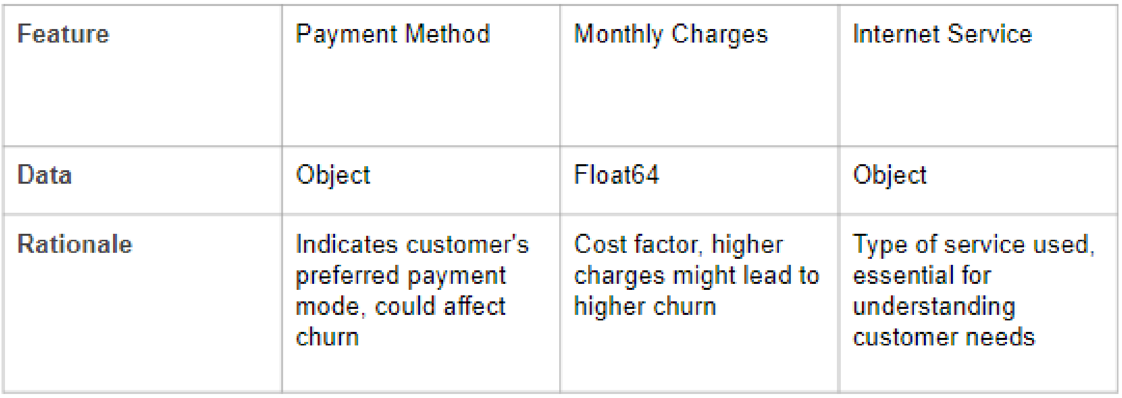
Variable Explanation

The features we used to determine the main contribution of churn were ‘Payment Method’ which would indicate customer’s preferred payment method, ‘Monthly Charges’ which was the cost factor and how higher charges might lead to higher churn, and ‘Internet Service’ which indicates the type of service used which is essential for understanding customer needs. We used them as the categorical feature and plotted them to understand the variables’ relationship to churn.



```
# Explore the distribution of selected categorical variables
categorical_features = ['Payment Method', 'Internet Service']
for col in categorical_features:
    plot_categorical_distribution(df, col, 'Churn')
```

1) Payment method - Payment method can show the preferred payment method chosen by each customer and provide important insight and their transaction behavior and preferences. Different types of payment methods as well as shifts in payment methods might include factors such as financial strain or dissatisfaction which would ultimately lead to churn.
2) Monthly Charges - Monthly charges essentially show how much customers are paying in a month-to-month basis. Higher monthly charges may lead to increased churn as customers who are cost-sensitive may look for more affordable alternatives. Being able to analyze the correlation between monthly charges and churn provides insight to how pricing and pricing structures impact customer retention. Through the analysis, being able to cater and optimize prices at the threshold where monthly charges significantly affect churn is crucial in meeting customer needs and expectations.
3) Internet Service - Internet services is a key determinant on a customer's specific needs and usage patterns. Being able to understand why a customer would opt into different plans such as basic or premium internet services to meet their needs is crucial as it can help optimize custom plans to meet customer needs. Catering to the varied requirements of their customer base in incredibly important in the world of customer retention.

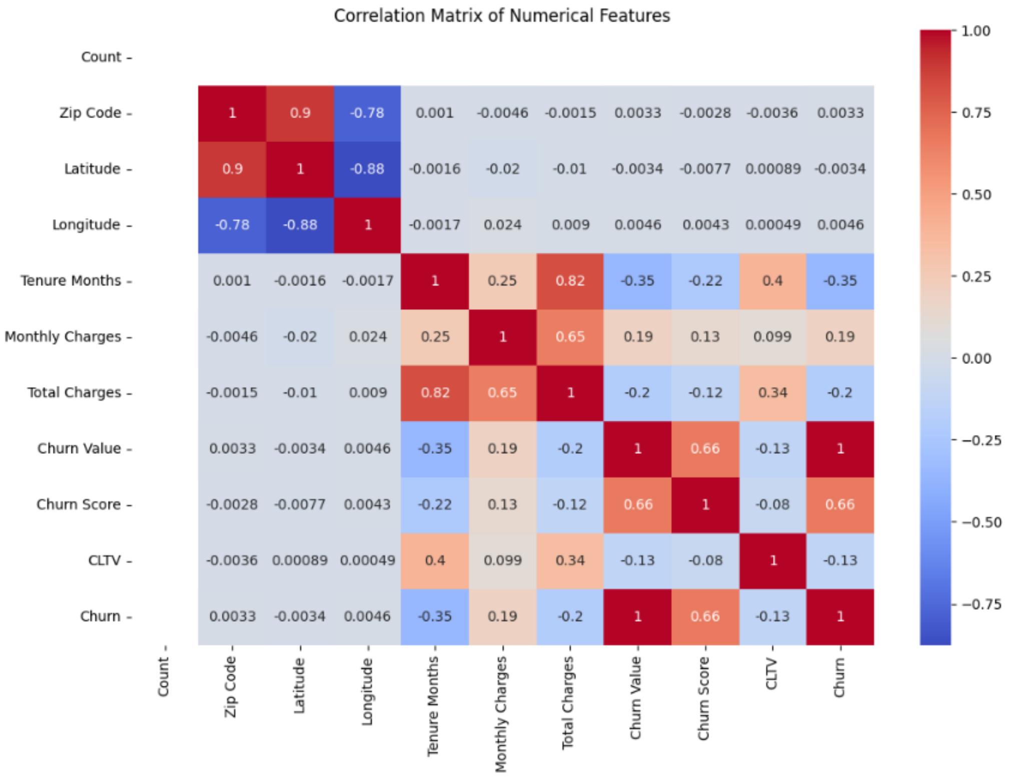


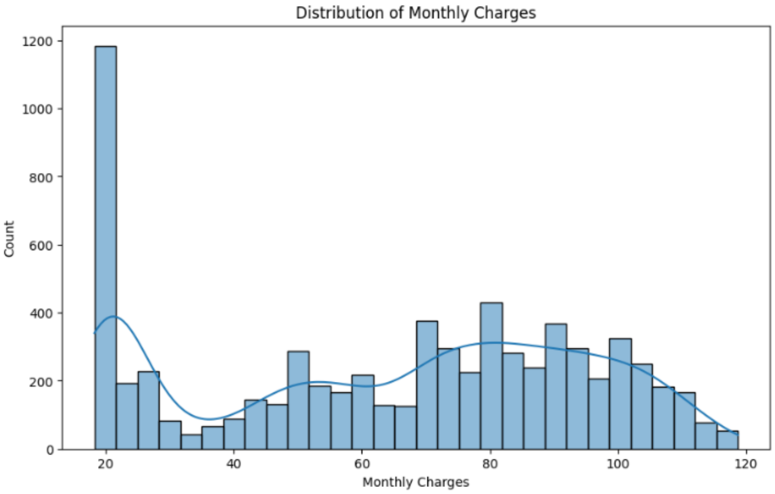
Feature	Payment Method	Monthly Charges	Internet Service
Data	Object	Float64	Object
Rationale	Indicates customer's preferred payment mode, could affect churn	Cost factor, higher charges might lead to higher churn	Type of service used, essential for understanding customer needs

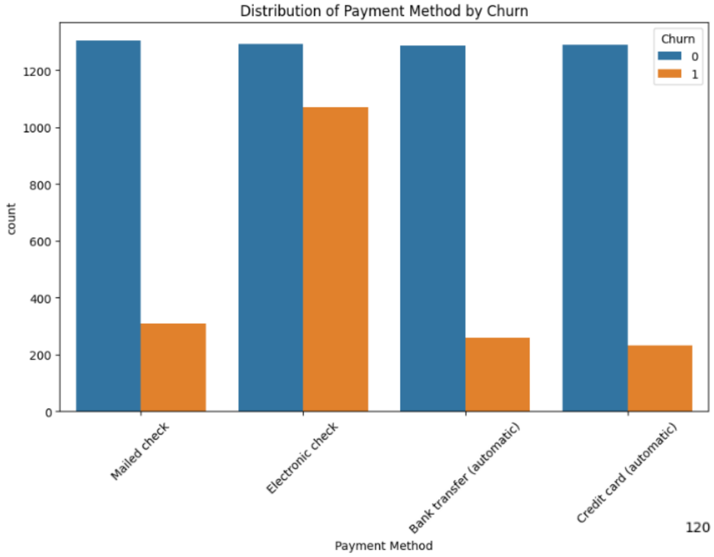
Methodology

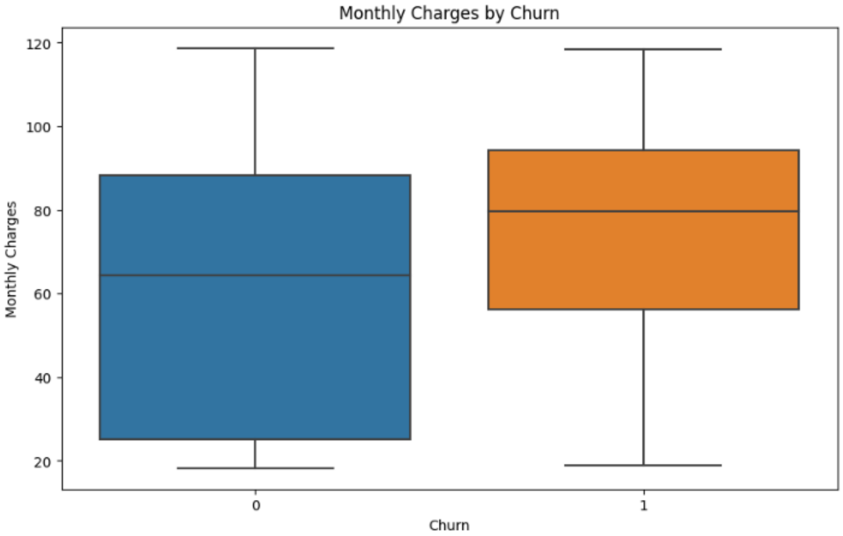
1) Exploratory Data Analysis (EDA)

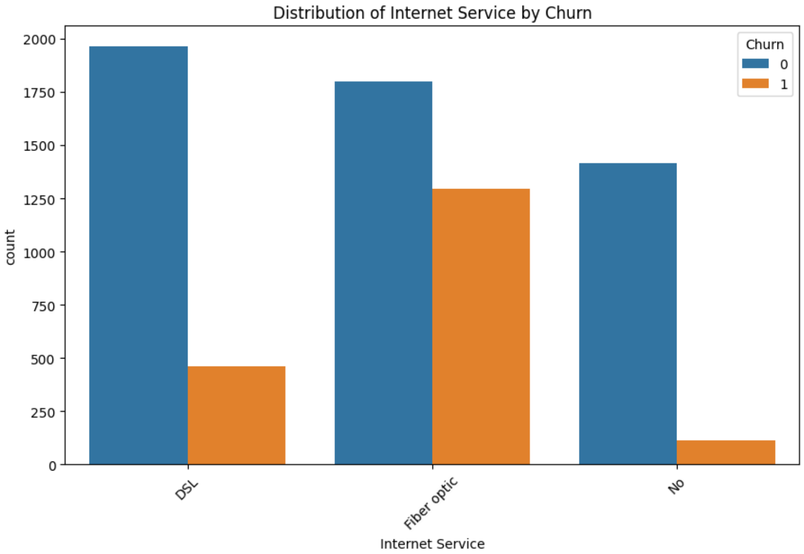
After opening up the dataset from kaggle.com using the pandas library, the libraries seaborn and matploblib.pyplot were used to make visualizations in the form of histograms, bar graphs, and box plots.




Distribution of Monthly Charges
Count
1200
1000
800
600
400
200
0
20
40
60
80
100
120
Monthly Charges


Distribution of Payment Method by Churn
count
1200
1000
800
600
400
200
0
Churn
0
1
Mailed check
Electronic check
Bank transfer (automatic)
Credit card (automatic)
Payment Method


Monthly Charges by Churn
Monthly Charges
120
100
80
60
40
20
0
1
Churn


Distribution of Internet Service by Churn
count
2000
1750
1500
1250
1000
750
500
250
0
Churn
0
1
DSL
Fiber optic
No
Internet Service

2) Resampling through SMOTE

After our exploratory data analysis, we noticed an imbalance in our dataset. SMOTE was used to create synthetic samples due to its ability to address the issue of class imbalance in our dataset.

EDA Results

Distribution of Payment Method by Churn

- Churn distribution demonstrate a larger number of customers not churning (0) to those who do (1).
- Majority of customers prefer an Electronic Check type of payment over other payment methods.
- Many Electronic Check users show a higher churn rate as well, suggesting that there is some sort of dissatisfaction or a need for improved payment processing.

Internet Service

- Fiber Optic users have the highest churn rate of all the other types of services, indicating some dissatisfaction with either the service provided or the price of the service itself.
- Customers with DSL service was shown to lower churn rates, possibly having something to do with perceived value or service satisfaction.

Monthly Charges

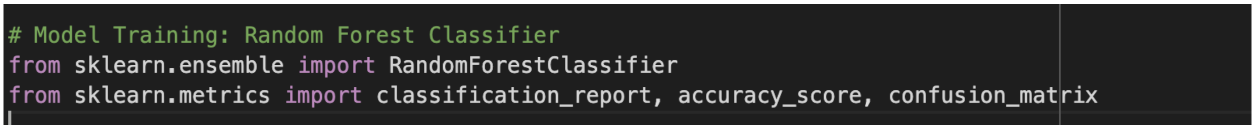
- Higher Monthly Charges correlate with increased churn, highlighting price sensitivity among customers.

SMOTE

In our analysis, we also used the Synthetic Minority Over-sampling Technique (SMOTE) to address the challenge of class imbalance, particularly in the 'Churn' classification. This technique, introduced by Chawla et al. in their 2002 paper, is pivotal for enhancing the performance of machine learning models in scenarios where the minority class is underrepresented (Chawla et al. 321-357). In our project, SMOTE was applied to the training dataset after the initial phase of preprocessing and feature engineering. By generating synthetic samples of the minority class, SMOTE significantly balanced the class distribution within the dataset. This balancing act was crucial for training the Random Forest Classifier. It ensured that the model did not show bias towards the majority class and improved its ability to generalize well to new unseen data. The effectiveness of this approach was evaluated on an untouched test set, where the model's performance metrics, including the accuracy and classification report, indicated a more robust and equitable handling of both classes.

Machine Learning Results

After the exploratory data analysis, the Random Forest Classifier was chosen for its ability to handle unbalanced datasets. The customer churn dataset suffered from class imbalances where other instances of one class significantly tends to outweigh the others. The classifier was trained on a balanced dataset achieved through the application of the Synthetic Minority Over-sampling Technique (SMOTE), ensuring that the minority class was adequately represented.



```
# Model Training: Random Forest Classifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, accuracy_score, confusion_matrix
```

Evaluation Metrics

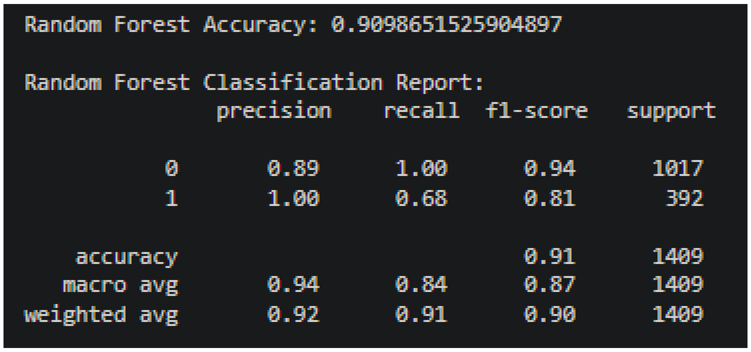
We trained our Random Forest model with hyperparameters to balance bias and variance, thereby optimizing the model's ability to generalize to unseen data. The model's hyperparameters, such as the number of trees in the forest ‘n_estimators’ and the maximum depth of the trees ‘max_depth’, were used to manage the complexity of the model and prevent overfitting.

After training, the model was evaluated using a test set that reflected the original distribution of the classes. Using this approach allowed us to assess the model's performance in a more realistic scenario, and ensured the validity and applicability of our findings.

Our evaluation metrics were precision, recall, and the F1-score, each providing insights into different aspects of the model's performance.

<u>Model Results</u>

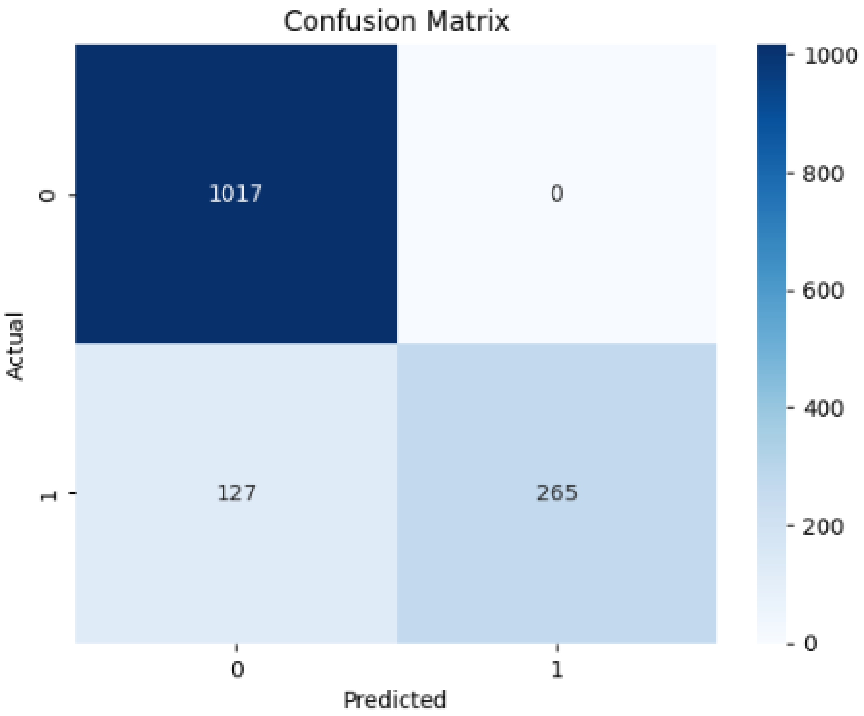
1) Precision, Recall, F1-Score:
    - Precision: For class 0 (no churn), the precision is 89%, indicates the model's accuracy in predicting customers who won't churn.
    - The model excels in recalling class 0 with a perfect score (100%), showing its effectiveness in creating a customer base who won't churn. While there is a perfect score for recalling class 0, there is a lower recall of 68% for class 1, suggesting different opportunities for identifying customers who do churn.
    - F1-Score: The F1-score, a balance between precision and recall, stands at 94% for class 0 and 81% for class 1, showcasing a good overall performance.



```
Random Forest Accuracy: 0.9098651525904897

Random Forest Classification Report:
              precision    recall  f1-score   support

           0       0.89      1.00      0.94      1017
           1       1.00      0.68      0.81       392

    accuracy                           0.91      1409
   macro avg       0.94      0.84      0.87      1409
weighted avg       0.92      0.91      0.90      1409
```

2) Confusion Matrix

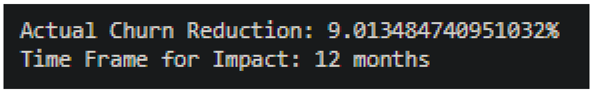
- A confusion matrix represents the prediction summary in matrix form and is an indicator for how many predictions are correct and incorrect per class.



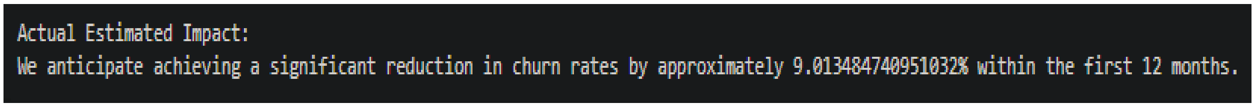
- True Positives (TP): 265 instances where the model correctly predicted churn.
- True Negatives (TN): 1017 instances where the model correctly predicted no churn.

- False Positives (FP): 0 instances where the model incorrectly predicted churn.
- False Negatives (FN): 127 instances where the model incorrectly predicted no churn.

This matrix was critical in visualizing the model's predictive power and diagnosing areas where the model could be improved. In particular, it was useful in reducing false negatives to capture more customers at risk of churning.



```
Actual Churn Reduction: 9.013484740951032%
Time Frame for Impact: 12 months
```



```
Actual Estimated Impact:
We anticipate achieving a significant reduction in churn rates by approximately 9.013484740951032% within the first 12 months.
```

Model Performance, Strengths and Areas for Improvement

Overall accuracy of the model stood at 90.99%, which is not only impressive, but it demonstrates the classifier's strong predictive capabilities. While accuracy can be used as a valuable metric, the additional context provided by precision, recall, and the F1-score is also essential for having a comprehensive understanding of the model's performance, particularly in an imbalanced class distribution scenario like churn prediction.

Some of the strengths of the model are demonstrated by its high precision in identifying non-churning customers and achieving perfect recall for this group. A high F1-score for class 0 indicated a well-balanced model for the majority class. However, the model displayed a relatively lower recall for class 1, which showed an area that needed to be improved. We would guess that increasing the recall for class 1 would mean enhancing the model's sensitivity to churn risk, thereby providing an opportunity to intervene and retain those customers.

In conclusion, implementing the Random Forest Classifier has worked in our favor. It has shown impressive capabilities in predicting customer churn. Its most notable strengths are its accuracy and its effectiveness in pinpointing true negative cases. However, there is room for improvement in enhancing its sensitivity towards instances of churn. These findings provide a foundation for subsequent model training phases, but they also provide insight. The groundwork for future iterations of model training should have a focus on improving recall for the churn class without compromising the overall accuracy and precision.

Proposed Solution

The insights gleaned from our machine learning model make it easy to reduce churn. We have proposed a solution that focuses on leveraging predictive analytics to tailor customer experiences and address the factors contributing to churn.

In terms of our strategic rationale, our Random Forest Classifier has provided us with a clear indication of which customers are at risk of churn. With this predictive power, we can adopt a proactive stance in customer relationship management. The strategic rationale for our solution is described by these points: customer segmentation, personalized engagement plans, and service enhancement initiatives.

For customer segmentation, we propose segmenting the customer base into distinct groups based on their risk of churn, service usage patterns, and billing information. This requires a good understanding of the unique preferences and behaviors of different customer segments which enables the creation of targeted marketing campaigns tailored to specific needs and concerns. This will ensure that interventions are timely and relevant to each customer's unique circumstances.

For our personalized engagement plans, we will be leveraging the segmentation to implement them. Customers identified with a higher risk of churn can be targeted with specialized campaigns, which may include personalized discount offers, service upgrades, or loyalty rewards. For those with high monthly charges or dissatisfaction with payment methods, we can offer customized billing solutions or payment plans to alleviate financial concerns.

For service enhancement initiatives, our data points to specific service types, such as Fiber Optic internet, correlating with higher churn rates. Addressing this, we propose service enhancement initiatives where we can refine the quality of service, adjust pricing strategies, or introduce new features that align with customer expectations and needs.

Estimated Impacts

The implementation of our solution is expected to achieve a significant reduction in churn rates, increase customer lifetime value, and enhance overall customer satisfaction. By aligning our efforts with predictive insights, we aim to foster a more loyal customer base, reduce acquisition costs, and bolster the company's market position.

Conclusion

The Telco Customer Churn project was quite the journey of data exploration, insightful analysis, and predictive modeling to confront the perennial challenge of customer churn. The synthesis of exploratory data analysis (EDA) and advanced machine learning techniques helped lead us to use a Random Forest Classifier, which would help identify customers at risk of discontinuing their services.

Our accuracy of 90.99% was impressive as it showed its effectiveness in distinguishing between customers who will churn and those who will remain. However, the machine learning model's recall for predicting churn (class 1) showed room for improvement. It is within this context that we made proposed solution to use targeted marketing campaigns and personalized customer engagement strategies to reduce churn. Our model's predictive strength informed us to use strategic customer segmentation, enabling personalized engagement plans and service enhancement initiatives tailored to individual customer profiles. We expect that this would shift the churn paradigm significantly.

The project's ending is not merely just the use of statistics but it is a testament to the power of data-driven decision-making. The anticipated churn reduction of 9.01% within the first 12 months is a powerful insight, as it can lead to new customer loyalty and satisfaction. As we advance, the project serves as a foundation for future innovations in predictive analytics and customer relationship management, showcasing a new era where customer retention can be redefined through data.

Appendix: Q & A Session

Following the presentation on our data science project, a brief Q&A session was held. This appendix documents the key question raised and the response provided.

Question 1: "Why did you use SMOTE in your analysis, given it was not covered in our course?"

Answer:

We decided to use SMOTE (Synthetic Minority Over-sampling Technique) in our analysis because we needed to address the issue of class imbalance in our dataset. We chose SMOTE for its effectiveness in generating synthetic samples for the minority class, thereby creating a more balanced dataset and improving the performance of our predictive models. Moreover, this approach ensures more accurate and equitable predictions across all classes, reflecting our commitment to addressing dataset biases and ensuring the reliability and fairness of our models.

Question 2: "Are you going to make a key feature table with the variables used?"

Answer: We said we would. Initially, our report and presentation did not include a key feature table. However, recognizing its importance for clarity and understanding, we committed to incorporating it in our final presentation and report. The key feature table we subsequently added provided a clear overview of the variables used, their descriptions, data types, and roles in our analysis, thereby enhancing the transparency and comprehensibility of our project.

Question 3: "You had too much text and not enough graphics in your presentation, are you going to use more graphics?"

Answer: We said we would. This feedback was crucial as we realized we were lacking in this regard of communication in data science. The necessary balance between text and visual elements. We significantly increased the use of visual elements, including charts, graphs, and infographics, to better illustrate our findings and methodology. These graphical enhancements not only made the presentation more engaging but also facilitated a clearer and more intuitive understanding of our analysis. The additional visuals provided a more dynamic and accessible way for the audience to grasp complex data and concepts, ensuring that our presentation was not only informative but also visually compelling.

References

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321-357.

Repositories and other links

https://github.com/leea36/I310DProjectCode.git

Open up the TelcoProject.ipynb and run the entire code from the beginning. Make sure to have the dataset in your drive to run the dataset. The dataset is here for convenience.

Telco Customer Project