Aarushi Jain (aj842) & Pooja Menon (pkm59)

Final Project Topic: Predicting Customer Retention Strategies Using Telco Customer Churn Dataset

Project Definition

The purpose of this project is to develop a predictive model that identifies customers who are most likely to stay with Telco. It looks at various attributes, such as customer demographics, service usage, and payment history to do so. The dataset that will be used for this project is on Kaggle and provides data that can be used to predict customer behavior to analyze customer retention. The goal is to use this data to come up with customer retention strategies to maintain valuable customers based on current behavior. Some other factors and questions that were asked about the data to identify customer loyalty and customer retention are the following:

1. What is the relationship between customer age, tenure, and churn?
2. How does gender influence the likelihood of churn?
3. Are customers with longer tenure less likely to churn?
4. How does the usage of specific services (e.g., Internet, StreamingTV, TechSupport) correlate with churn?
5. Do customers who use more services (e.g., multiple services like InternetService, TechSupport, StreamingTV) have a lower churn rate?
6. Do customers who use paperless billing have a higher retention rate?
7. What are the key factors contributing to customer churn?
8. How does the length of the customer relationship (tenure) correlate with churn?
9. Are customers with contract types that lock them in (e.g., year-long contract) less likely to churn than those with month-to-month contracts?
10. Is there a correlation between the type of internet service and customer churn?
11. How does churn differ between customers with and without internet service?
12. What is the churn rate for customers who have only one service versus those who have multiple services (e.g., Internet + Tech Support)?
13. What are the retention rates based on the number of services a customer subscribes to?
14. How does customer satisfaction with the service (inferred from usage patterns) affect retention?
15. What behaviors or attributes can predict early churn (e.g., a decline in service usage over time)?

Telecom companies face a lot of challenges in retaining customers. Predictive models can be used to help identify customers who are likely to remain loyal. By predicting this customer loyalty, the company can allocate resources more effectively and efficiently to offer personalized efforts to these customers and improve their overall customer loyalty. The strategic aspects of the project will include managing customer data, cleaning the data for machine learning, and optimizing retention programs through model creation and analysis. The project applies concepts that we learned in lecture, such as data cleaning, feature engineering, and creating models. We will use various models to predict customer retention and evaluate their performance in terms of accuracy and effectiveness.

Novelty & Importance

This project and customer retention is important as it is more cost-effective for the company rather than having to acquire new customers. By predicting the customers that are likely to stay, Telco can design personalized marketing and customer service strategies that will help in building long-term customer loyalty. The ability to predict retention can also help optimize the company's efforts in customer satisfaction. Additionally, this project allows us to explore customer behaviors that lead to loyalty and engagement. The key challenges in current data management practices for this project are dealing with missing values, handling categorical features effectively, and understanding the relationships between multiple attributes. The dataset that we have chosen also requires careful balancing to avoid overfitting, as customer retention might be influenced by several factors that need to be modeled accurately.

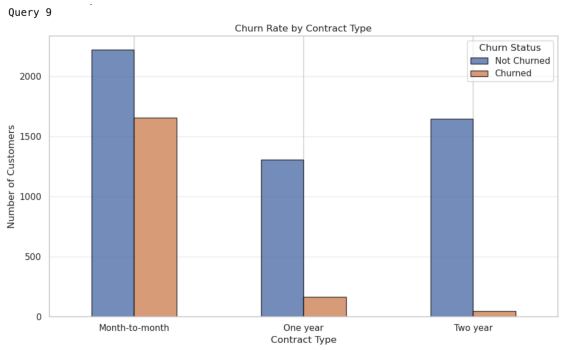
Progress & Contribution

The dataset that was used for this project has been derived from Kaggle and is titled Telco Customer Churn. It contains customer information from a telecommunications company, with features such as gender, the type of service, tenure, their payment history, and whether or not the customer churned. After being downloaded as a csv file from the Kaggle website, the data was cleaned to either fill or remove missing values depending on how unique they are based on mean, median, and mode, to transform the categorical variables to numerical ones, and to also remove certain columns if necessary.

In order to answer the question and solve the problem of predicting the likelihood of customer churn, we created multiple Machine Learning models, which includes Logistic Regression and Random Forest Classifier. We were able to train these models on a training dataset and evaluated them on a testing dataset. In order to measure its effectiveness, we compared their performance based on accuracy, precision, recall, F1-score, and cross-validation results. The F1-score and cross-validation results of a model prove to be important as they provide insights into how well the model is performing. More specifically, the F1-score is a measure of a model's accuracy that combines both precision and recall into a single metric. It is especially useful when dealing with class imbalances where the number of customers who churn might be much smaller than the number who do not churn. On the other hand, cross-validation is a technique that is used to assess how well a model generalizes to unseen data. It works by splitting the dataset into multiple subsets, training the model on those subsets, and then testing it on the remaining ones. If a model performs consistently well across the different folds, it indicates that it is generalizing well and is not overly sensitive to the particular partition of the data used. Conversely, if there is high variability in performance between folds, it suggests that the model might not be robust or could be overfitting, which is what the models should avoid. The goal of doing this was to identify which one would be the most effective model for predicting customer churn. Additionally, we were able to use SQL to look at the different features present in the dataset and how one factor might have affected the other. This type of querying helped us look at the data more in depth by writing queries to answer the questions listed above and draw connections to the churn rate. Listed below are some of the most important features that we found to help us predict customer churn and help us to show customer loyalty.

1. Question 9: Relationship Between Contract Type & Customer Churn

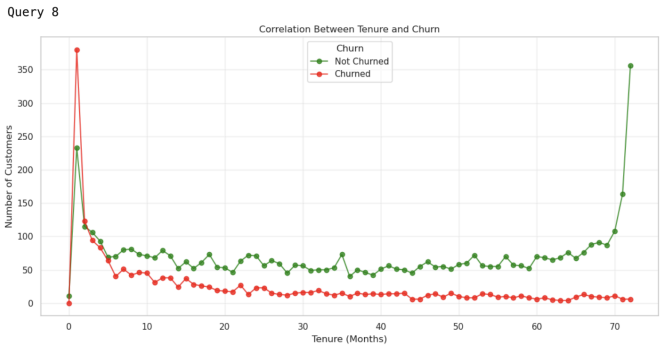
	Contract	Churn	num_customers
0	Month-to-month	0	2220
1	Month-to-month	1	1655
2	One year	0	1307
3	One year	1	166
4	Two year	0	1647
5	Two year	1	48



Based on the output above, it is clear that month-to-month contracts have significantly higher churn compared to one-year and two-year contracts.
Month-to-month contracts allow more flexibility for customers to leave, while longer-term contracts may bind customers and encourage retention through incentives like discounts. This insight suggests focusing retention strategies on month-to-month contract customers, perhaps by offering benefits for upgrading to a longer-term contract.

2. Question 8: Relationship Between Tenure & Churn

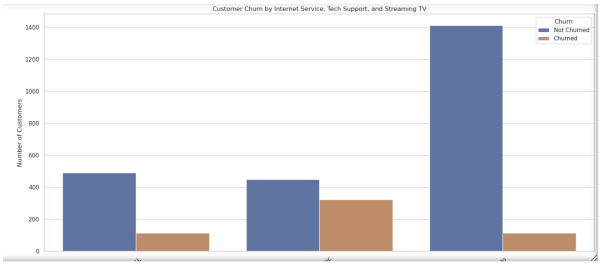
	tenure	Churn	num_customers
0	0	0	11
1	1	0	233
2	1	1	380
3	2	0	115
4	2	1	123
..	...	...	...
140	70	1	11
141	71	0	164
142	71	1	6
143	72	0	356
144	72	1	6



Customers with short tenure seem to show a higher churn rate compared to customers with longer tenure. You can also see that as tenure increases, churn decreases. This can be seen if you look at the correlation between the tenure and num_customers column. This indicates that the length of the customer relationship strongly correlates with churn and that new customers are at a higher risk of churning.

3. Question 5: Relationship Between Internet Service and Specific Service Usage

	InternetService	TechSupport	StreamingTV	Churn	num_customers
0	DSL	No	No	0	619
1	DSL	No	No	1	261
2	DSL	No	Yes	0	279
3	DSL	No	Yes	1	84
4	DSL	Yes	No	0	513
5	DSL	Yes	No	1	71
6	DSL	Yes	Yes	0	551
7	DSL	Yes	Yes	1	43
8	Fiber optic	No	No	0	544
9	Fiber optic	No	No	1	560
10	Fiber optic	No	Yes	0	585
11	Fiber optic	No	Yes	1	541
12	Fiber optic	Yes	No	0	192
13	Fiber optic	Yes	No	1	50
14	Fiber optic	Yes	Yes	0	478
15	Fiber optic	Yes	Yes	1	146
16	No	No internet service	No internet service	0	1413
17	No	No internet service	No internet service	1	113

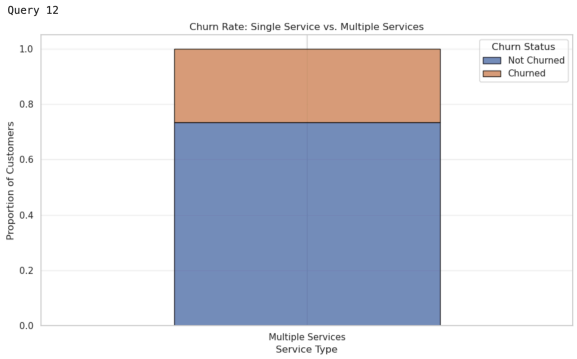


Customers using the fiber optic internet service have a higher churn rate compared to DSL, and those without internet service have the lowest churn rate. Fiber optic users may face higher prices or technical issues leading to higher churn. DSL and non-internet users are possibly more budget-conscious, which is why they exhibit lower churn. This suggests investigating Fiber optic service satisfaction and pricing to reduce churn in this segment.

4. Question 12: Relationship Between Single vs. Multiple Services

	Service_Type	Churn	num_customers
0	Multiple Services	0	5174
1	Multiple Services	1	1869

Query 12

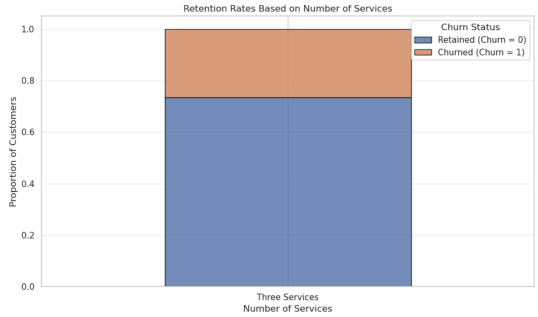


Customers with multiple services churn less compared to those with fewer services. This means that customers using multiple services may have stronger engagement or dependency, reducing churn. Bundling services could be an effective strategy to boost loyalty and reduce churn among single-service users.

5. Question 13: Relationship Between Retention and Number of Services

	Service_Type	Churn	num_customers
0	Three Services	0	5174
1	Three Services	1	1869

Query 13

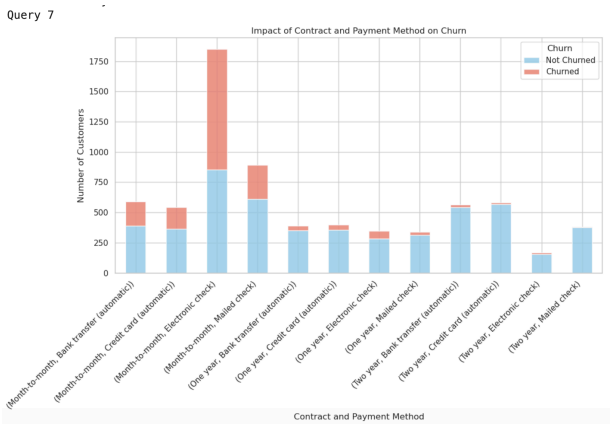


Customers with three services churn less than those with fewer services. This suggests that increasing the number of services a customer subscribes to strengthens retention. Cross-selling and promoting service bundles could significantly improve customer loyalty.

6. Question 7: Relationship Between Payment Method, Contract Type, and Churn

	Contract	PaymentMethod	Churn	num_customers
0	Month-to-month	Bank transfer (automatic)	0	388
1	Month-to-month	Bank transfer (automatic)	1	201
2	Month-to-month	Credit card (automatic)	0	365
3	Month-to-month	Credit card (automatic)	1	178
4	Month-to-month	Electronic check	0	856
5	Month-to-month	Electronic check	1	994
6	Month-to-month	Mailed check	0	611
7	Month-to-month	Mailed check	1	282
8	One year	Bank transfer (automatic)	0	353
9	One year	Bank transfer (automatic)	1	38
10	One year	Credit card (automatic)	0	357
11	One year	Credit card (automatic)	1	41
12	One year	Electronic check	0	283
13	One year	Electronic check	1	64
14	One year	Mailed check	0	314
15	One year	Mailed check	1	23
16	Two year	Bank transfer (automatic)	0	545
17	Two year	Bank transfer (automatic)	1	19
18	Two year	Credit card (automatic)	0	568
19	Two year	Credit card (automatic)	1	13
20	Two year	Electronic check	0	155
21	Two year	Electronic check	1	13
22	Two year	Mailed check	0	379
23	Two year	Mailed check	1	3

Query 7



For month-to-month contracts, churn is highest among customers using electronic checks, followed by mailed checks. Customers with longer-term contracts and automatic payments have much lower churn rates. Electronic and mailed check users may face friction in payment processes, leading to dissatisfaction and higher churn. Encouraging automatic payments and offering incentives to month-to-month customers could reduce churn.

7. **Question 6: Relationship Between Paperless Billing and Churn**

	PaperlessBilling	Churn	num_customers
0	0	0	2403
1	0	1	469
2	1	0	2771
3	1	1	1400

Query 6

Paperless billing users have higher churn rates than those not using it. While paperless billing is convenient, it might be less effective at keeping customers engaged or aware of their payments. Adding reminders or promotional offers could help mitigate churn.

Looking at these features and asking these questions are important to addressing the problem of customer churn because they help identify patterns, behaviors, and characteristics that influence a customer's likelihood to leave. By analyzing factors such as contract type, tenure, service usage, and payment methods, we can pinpoint high-risk customer groups and the underlying reasons for why they have a higher rate of leaving. For instance, understanding that month-to-month customers have higher churn rates or that tenure correlates inversely with churn rates will allow companies to target their retention strategies towards those customers. They can offer incentives for customers to take on longer-term contracts or they can enhance early customer engagement.

Advantages & Limitations

There are many advantages in the approach that we take for this project. First, we are able to look at multiple factors that could affect customer churn by exploring different customer attributes. By asking questions related to many of the features provided in the kaggle dataset and performing data analysis for each of them, we are able to consider both demographic and usage-based aspects. It allows us to understand the behavior of our customers and also identify the different customer segments that we are working with to help Telecom companies. Additionally, the use of data visualization is another advantage as we are able to actually see the trends that we are looking for. It provides a way for us to observe the relationship between different variables. We are also able to pinpoint certain activities that we see to the specific features that we are working with. Along with this, our approach with our machine learning models further this analysis by providing predictive insights and allowing us to forecast customer churn.

However, there could be limitations that arise with our approach. Even though we cleaned our data before performing our analysis, there could be some issues with the quality of our data. For instance, there could be some missing values that can still have an impact on the results of our analysis or lead to inaccurate predictions. There are non-numeric values in multiple columns, such as TotalCharges, which could affect this. Our machine learning models also depend heavily

on the quality of the data that we present, which means that our predictions could be impacted as well.

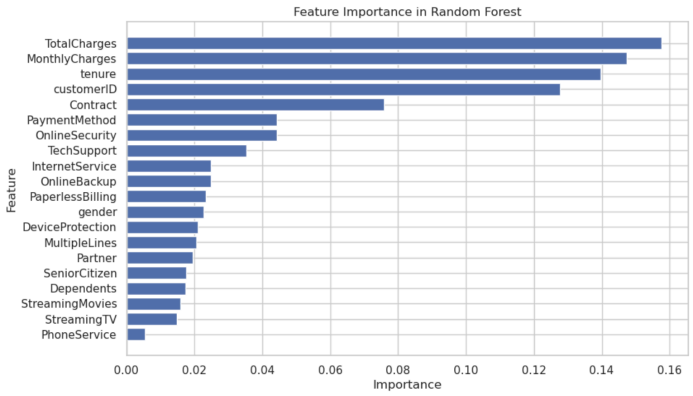
Model Creation & Analysis of the Model

Output of Random Forest Classifier Model

```
Accuracy: 0.7955997161107168
Confusion Matrix:
 [[939  97]
 [191 182]]
Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.91      0.87      1036
           1       0.65      0.49      0.56       373

    accuracy                           0.80      1409
   macro avg       0.74      0.70      0.71      1409
weighted avg       0.78      0.80      0.79      1409
```



The accuracy, according to the output given by the model, shows that the model is correctly predicting class labels for 79.56% of the data. This is a sizable amount, meaning the model performs reasonably well. The graph on the right demonstrates feature importance in the random forest model, with the most important features being TotalCharges, MonthlyCharges, and tenure. This means that these three factors most influence Telco customer churn. In the classification report, it is seen that 83% of the instances predicted to be “0” are actually “0” for the “0” class, while the 91% recall rate shows that 91% of the “0” instances are correctly identified.

Output of Logistic Regression Model

```
Confusion Matrix: [[935 101]
 [159 214]]
              precision    recall  f1-score   support

           0       0.85      0.90      0.88      1036
           1       0.68      0.57      0.62       373

    accuracy                           0.82      1409
   macro avg       0.77      0.74      0.75      1409
weighted avg       0.81      0.82      0.81      1409

Accuracy: 0.815471965933286
```

The logistic regression model shows that the model works pretty well with an 81.55% rate of accuracy. The macro average row in both of the above models shows the unweighted average of precision, recall, and f1-score for both classes. The confusion matrix shows the number of true positives, true negatives, false positives, and false negatives – 935, 101, 159, and 214 respectively.