Proposal Page

Midpoint Page

Saumya Bajaj, Katniss Min, Liane Nguyen, Calvin Truong, Xiangyi Zhu

Final Report

Introduction and Background

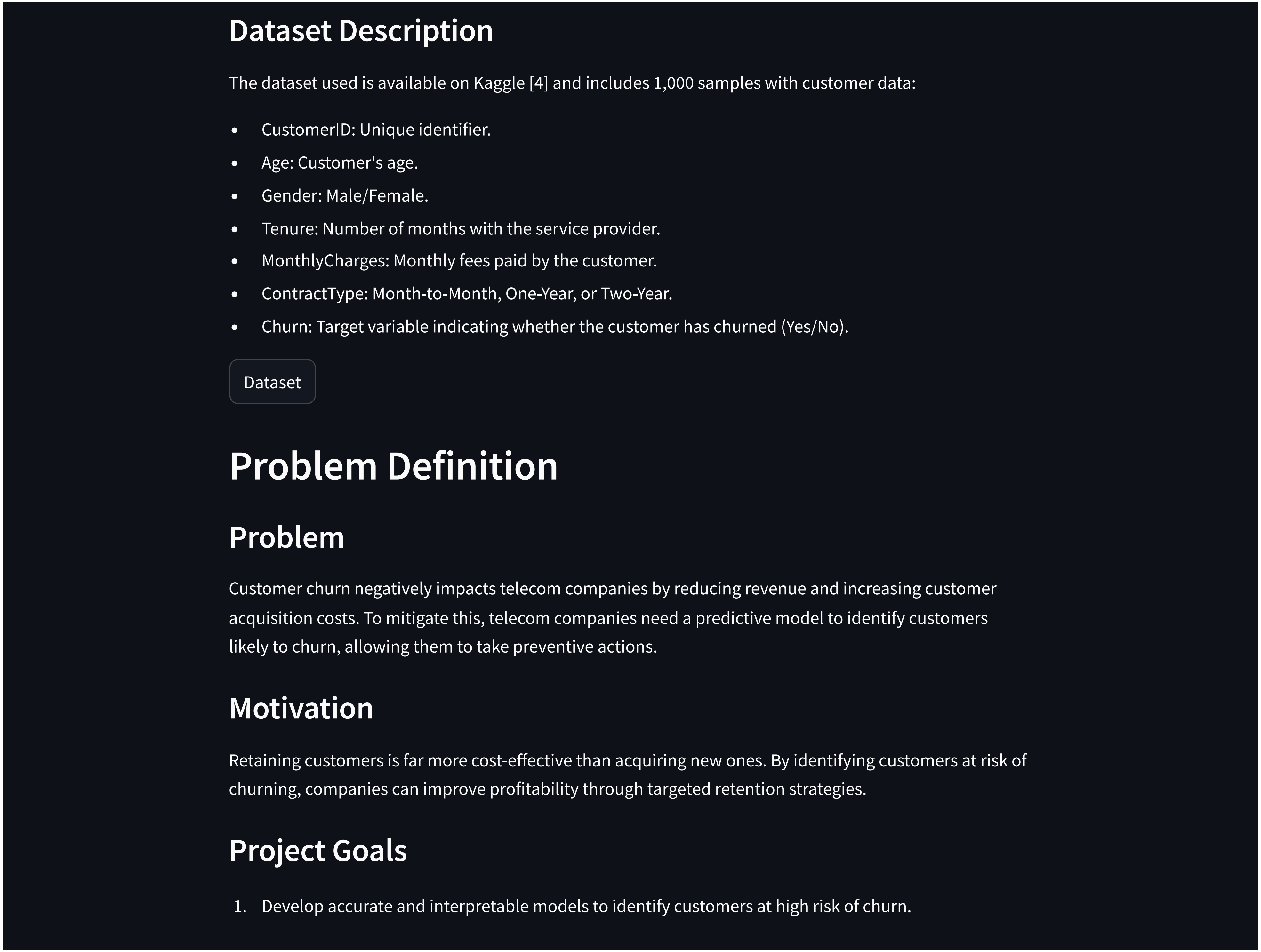
Introduction

Customer churn is an important issue in the telecom industry, leading to revenue loss and additional costs to acquire new customers. To tackle this, companies have started predicting customer churn by identifying those at risk. This project will aim to develop a machine learning model that accurately predicts customer churn, minimizing profit loss and improving customer retention.

Literature Review

The majority of the literature aims to predict whether a customer will churn. J. Bhattacharyya and M. K. Dash identified “ten overarching groups of scholarship” [1], the first two being churn prediction and modeling and feature selection techniques and comparison. Another paper proposes a six-step methodology starting with data pre-processing and feature analysis [2]. However, despite the extensive research and demand for churn reduction tools, there is a lack of “well-defined guidelines on appropriate model evaluation measures” [3].

There is a great demand for cross-industry model evaluation as well as a more generalized churn prediction model, as most are performed on one specific industry or consumer base [1]-[3].



Dataset Description

The dataset used is available on Kaggle [4] and includes 1,000 samples with customer data:

- CustomerID: Unique identifier.
- Age: Customer's age.
- Gender: Male/Female.
- Tenure: Number of months with the service provider.
- MonthlyCharges: Monthly fees paid by the customer.
- ContractType: Month-to-Month, One-Year, or Two-Year.
- Churn: Target variable indicating whether the customer has churned (Yes/No).

Problem Definition

Problem

Customer churn negatively impacts telecom companies by reducing revenue and increasing customer acquisition costs. To mitigate this, telecom companies need a predictive model to identify customers likely to churn, allowing them to take preventive actions.

Motivation

Retaining customers is far more cost-effective than acquiring new ones. By identifying customers at risk of churning, companies can improve profitability through targeted retention strategies.

Project Goals

1. Develop accurate and interpretable models to identify customers at high risk of churn.

2. Use PCA to simplify the dataset and highlight the most significant customer features for actionable insights.
3. Promote sustainability by reducing churn, minimizing resource-intensive customer acquisition.
4. Avoid bias in predictions, ensuring fair treatment of sensitive demographic features.

Methods

Data Preprocessing Methods

- One-Hot Encoding: We used one-hot encoding to transform categorical variables into binary columns, making the data suitable for machine learning models. This was especially necessary for features like Gender, ContractType, and InternetService, which require numerical representation.
- MinMax Scaling: MinMax scaling was used to normalize continuous variables, ensuring that features like MonthlyCharges and Tenure have comparable ranges, which aids in model performance.
- Imputation: Missing values were imputed to avoid gaps in the data, ensuring completeness for analysis.

Machine Learning Models

1. Principal Component Analysis (PCA) (Unsupervised):

- Goal: PCA aims to reduce dataset dimensionality while preserving key variance, identifying the most critical components driving customer behavior. This simplification minimizes redundancy and improves efficiency for further analysis, reducing the risk of feature overlap and distortion.
- Effectiveness: PCA retained 91.18% of the dataset's variance using only two components, capturing essential information while reducing feature redundancy. This streamlined feature space enhances computational efficiency, especially for clustering models like K-Means, and supports logistic regression by providing a clear, two-dimensional visualization that aids in distinguishing churned from non-churned customers.

2. Logistic Regression (Supervised):

- Goal: Logistic regression is designed to predict customer churn probability by analyzing demographic and contract features, focusing on binary classification (churn vs. not churn). This approach is well-suited for identifying at-risk customers, supporting proactive retention efforts.
- Effectiveness: Logistic regression efficiently handles large datasets, providing interpretable coefficients that reveal each feature's impact on churn. This interpretability supports targeted retention strategies, and the model's linearity allows for quick, effective classification, making it ideal for timely churn prediction and enabling telecom companies to make data-driven retention decisions.

3. K-Means Clustering (Unsupervised):

- Goal: Group customers into distinct clusters based on demographic and service usage patterns to identify high-risk customer segments. This approach provides the features of high risk churning customers for pattern analysis, and also prediction for behaviors and use cases.
- Effectiveness: K-Means is able to identify similar patterns of customers who churned, and it can create new groups of clustered data which could be feasible for prediction training and churning pattern analysis.

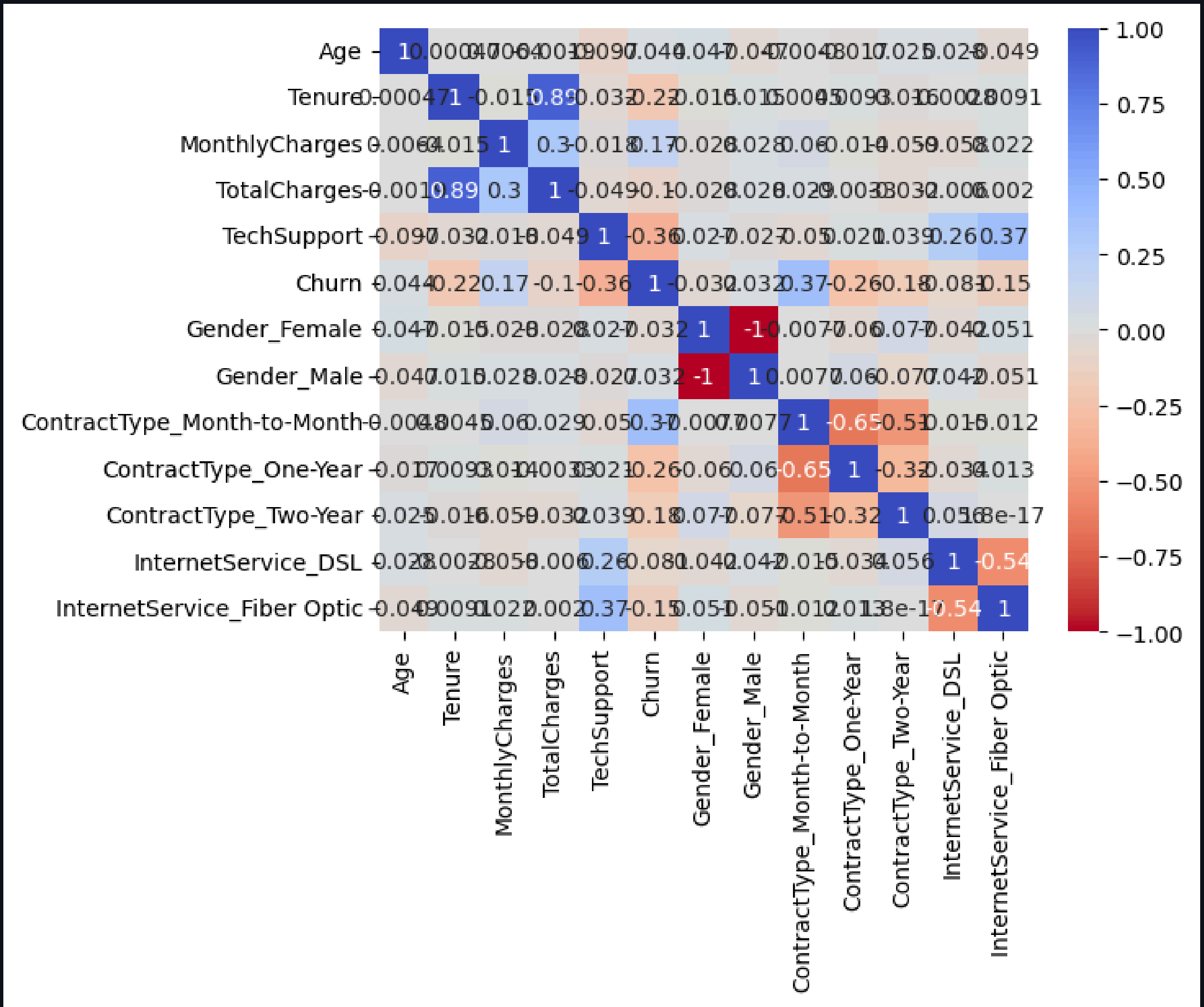
Results and Discussion

Data Preprocessing Method

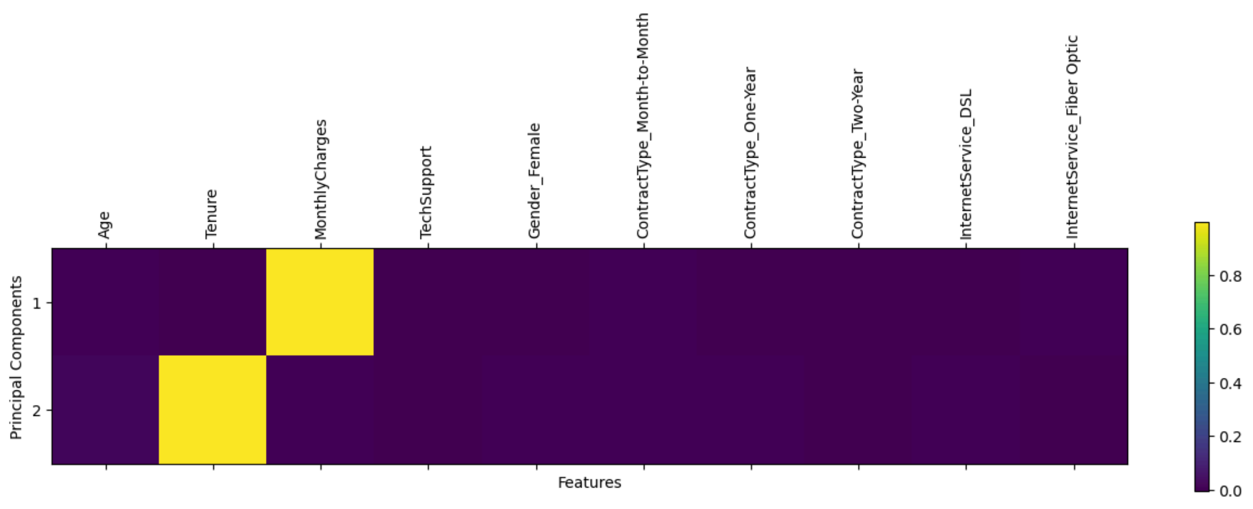
- One-Hot Encoding
  - Before applying PCA, we used one-hot encoding to convert categorical features (Gender, ContractType, and InternetService) into numerical form. This transformation was essential because machine learning models, including PCA, require numerical inputs to understand patterns effectively. One-hot encoding created new binary columns representing each category (e.g., Gender_Female, Gender_Male), allowing the model to differentiate categories without assuming any ordinal relationship. This approach ensured that features with categorical values could be meaningfully included in our analysis.

PCA Model Visualizations

- Correlation Matrix
    - The correlation matrix heatmap displays how certain features relate to each other, indicating which are strongly correlated. This was essential in deciding to use PCA because reducing these correlations simplifies the dataset, making the model more efficient and focused. In our case, we found high correlations with features like TotalCharges and Gender_Male, so we dropped them to prevent redundancy. This step supports PCA by ensuring only the most meaningful features remain.



- Principal Components Map
  - This visualization shows how each feature contributes to the two main principal components derived by PCA. By examining this heatmap, we see the features that strongly influence each component, helping us understand what drives patterns in customer churn. For instance, ContractType and InternetService may show strong contributions, highlighting their importance in customer behavior. This visualization also assists in interpreting the role each feature plays after reduction, even though PCA combines them into components.



PCA Model Qualitative Metrics

- Explained variance
  - PCA captured around 91.18% of the total variance using only two components. This means we retained most of the important information from the original data in a simpler form, reducing the number of dimensions without sacrificing key patterns. The high percentage (91.18%) indicates that our dataset was simplified while still holding onto the main insights about customer behavior, enhancing efficiency and accuracy for downstream models like logistic regression.

PCA Model Analysis

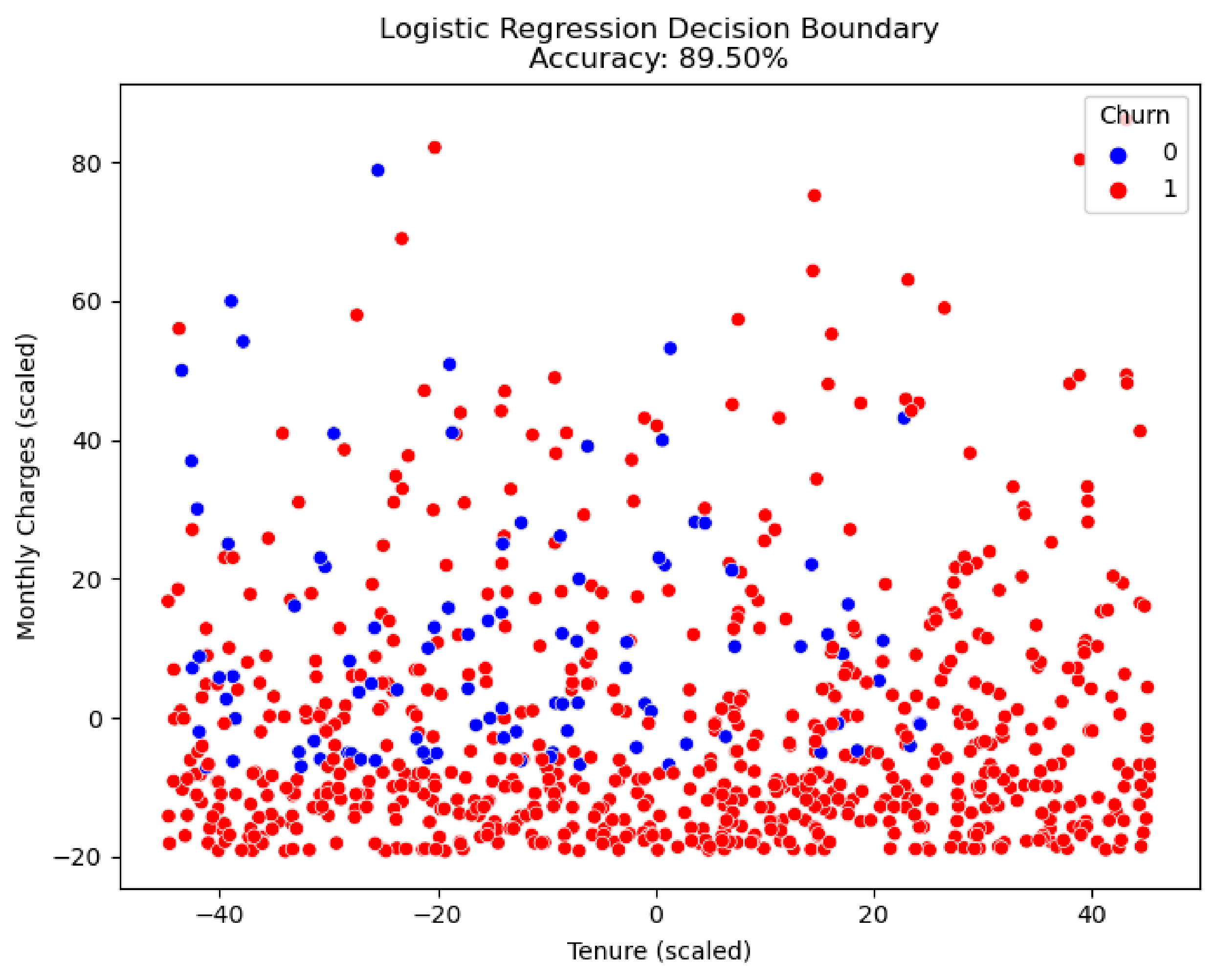
The PCA transformation worked well because it kept a large portion of the original data’s variance. This simplification helped the Logistic Regression model focus on the most informative parts of the dataset, removing noise and redundant features that could complicate analysis. Overall, the PCA method improved classification by providing clear, concise feature representations. By reducing feature overlap, PCA made the dataset easier to interpret and improved the stability of Logistic Regression results. Also, reducing the data to two main components allowed us to create visualizations that reveal decision boundaries in the Logistic Regression model, helping us see how well the model distinguishes between churned and non-churned customers. However with PCA, the features are combined into components, which can make it harder to interpret specific impacts of each feature on churn. This trade-off is typical in dimensionality reduction methods, where some detail is lost in favor of simplicity.

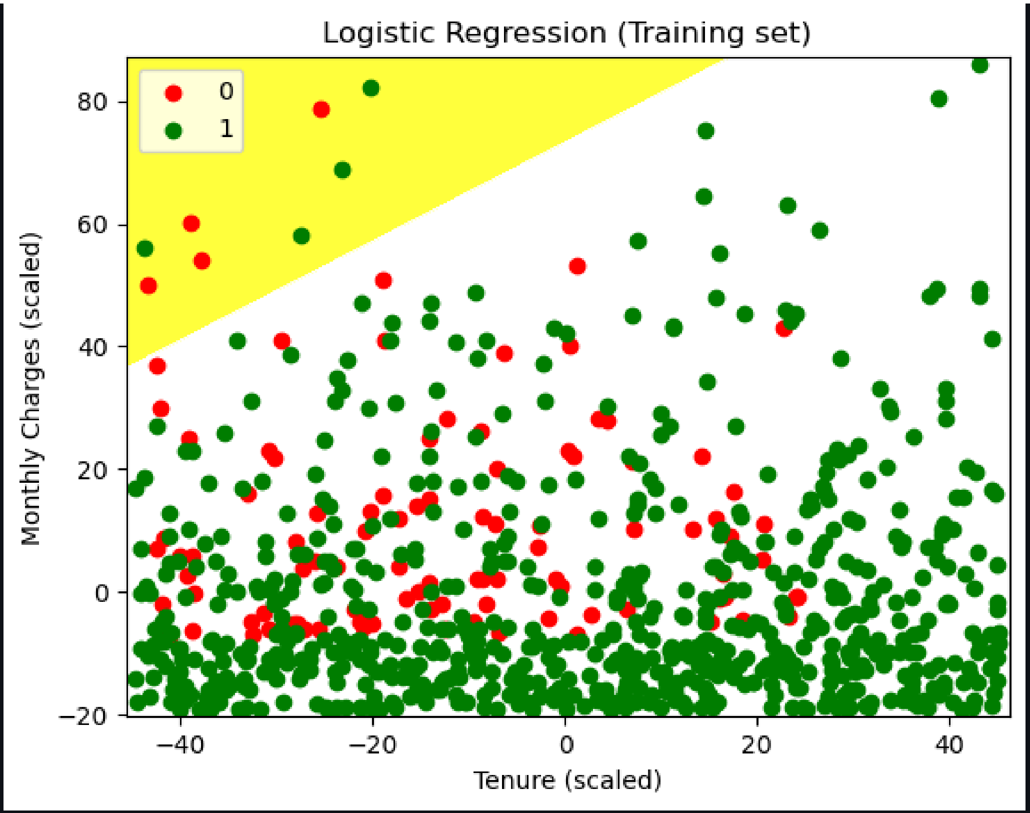
PCA Model Next Steps

- Consider testing three components instead of two to see if this slightly improves the explained variance while keeping computations efficient. This could enhance the model’s ability to capture more subtle patterns without overloading it.
- Experiment with other feature scaling or transformation methods before PCA to see if they help capture variance more effectively. Small adjustments here may help the Logistic Regression model work even better.
- We can test different thresholds for variance retention in PCA (for example, setting it at 95% instead of 91%) to find an optimal balance between simplicity and accuracy. Higher retention may increase the logistic regression’s ability to classify churn without overfitting.

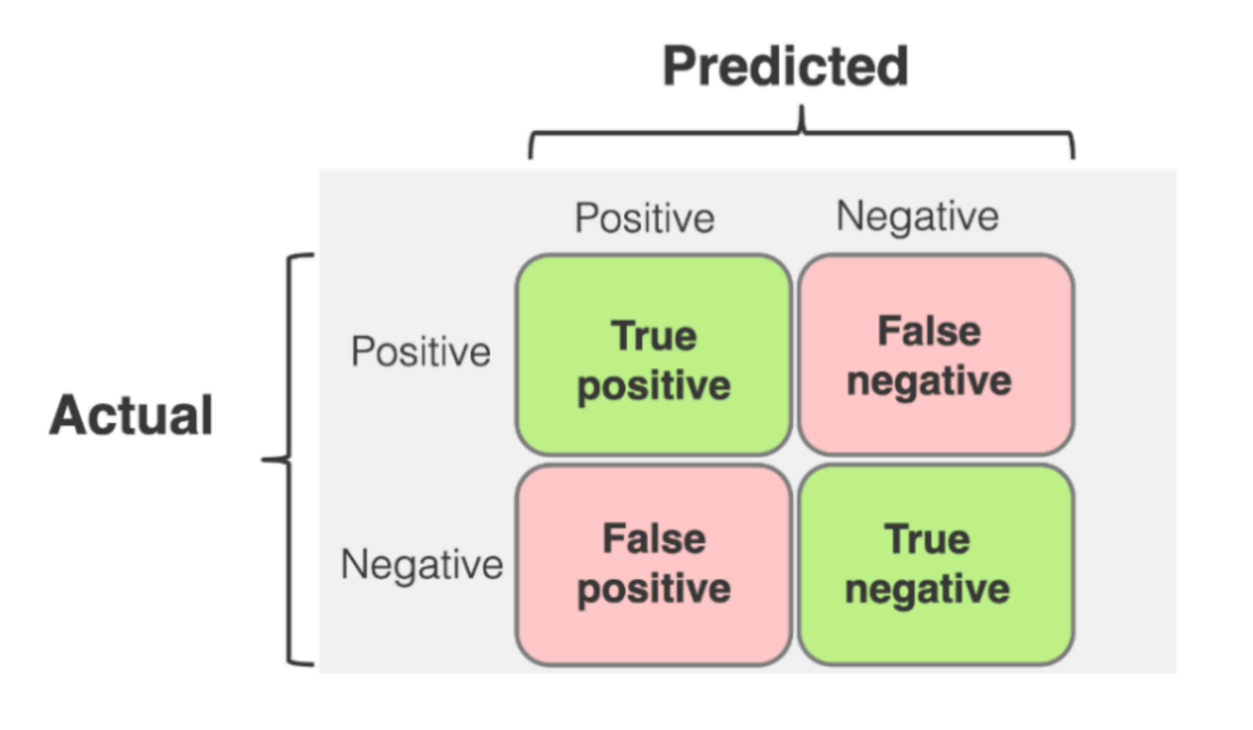
Logistic Regression Model Visualizations

- Scatter Plot with Decision Boundary:
    - To visualize the logistic regression model’s effectiveness in separating churned and non-churned customers, we created a scatter plot showing the decision boundary in the transformed PCA space. This plot allowed us to observe how well the model differentiates between customer classes within a simplified, two-component feature space.

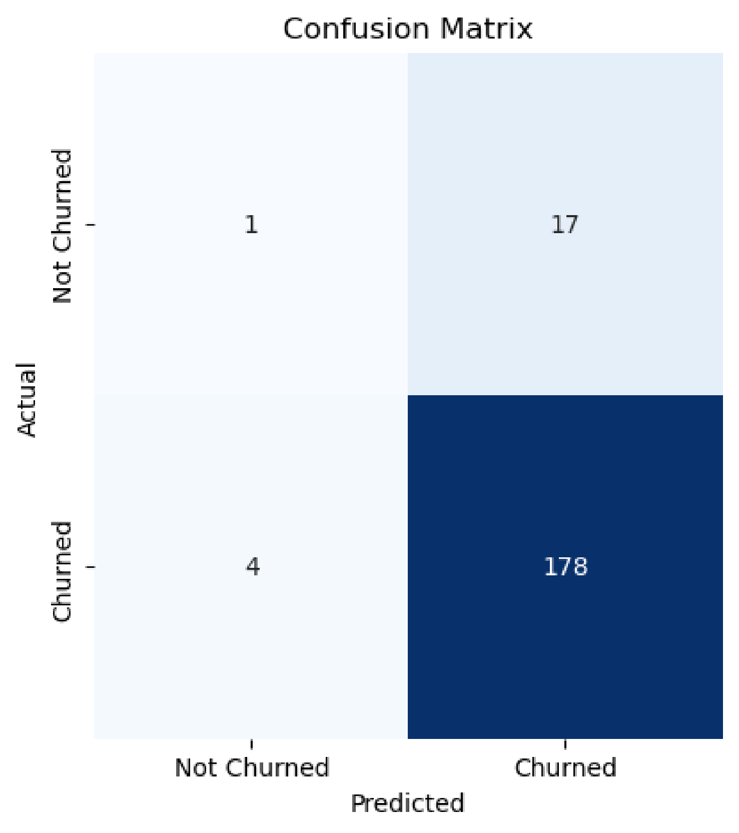

Logistic Regression Decision Boundary
Accuracy: 89.50%
Churn
0
1
Monthly Charges (scaled)
80
60
40
20
0
−20
−40
−20
0
20
40
Tenure (scaled)



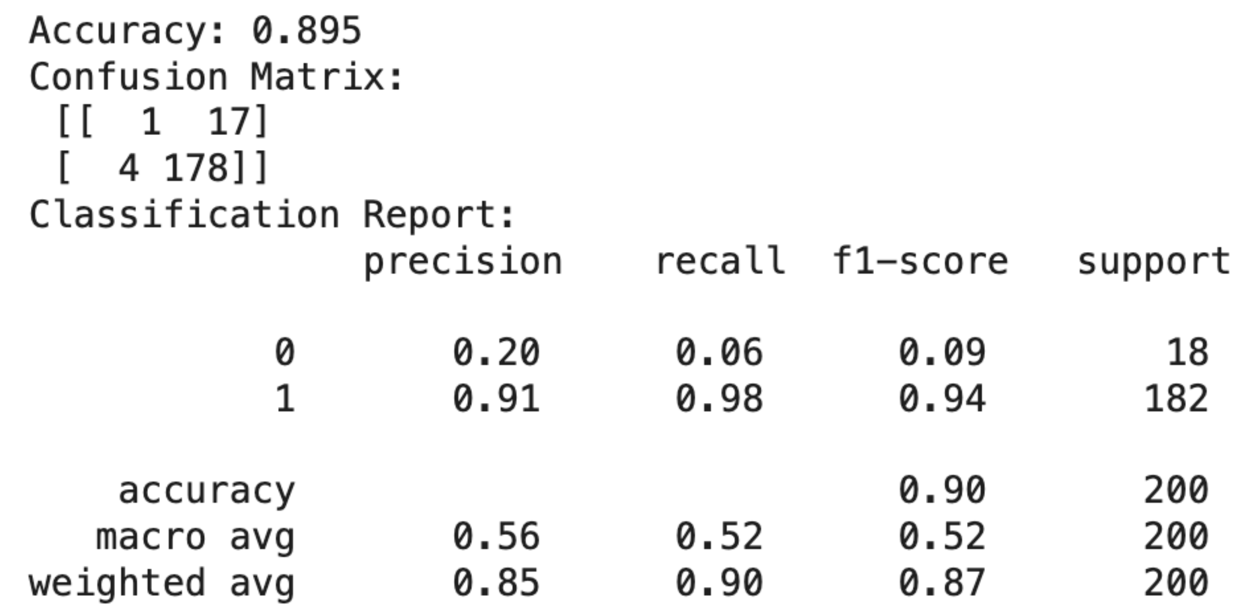
- Confusion Matrix:
  - The confusion matrix illustrates the model's performance on the test data, showing counts for true positives (churned correctly identified), true negatives (not churned correctly identified), false positives, and false negatives. This visualization helps evaluate the model's accuracy and the distribution of prediction errors, revealing strengths and areas for improvement.



		Predicted	
		Positive	Negative
Actual	Positive	True positive	False negative
	Negative	False positive	True negative



- Classification Report:
  - The classification report includes precision, recall, F1-score, and support metrics for each class, with a focus on the “churned” category. High recall and F1-score for the churned class indicate the model’s strong ability to detect at-risk customers, which is essential for customer retention.



```
Accuracy: 0.895
Confusion Matrix:
 [[  1  17]
 [  4 178]]
Classification Report:
              precision    recall  f1-score   support

           0       0.20      0.06      0.09        18
           1       0.91      0.98      0.94       182

    accuracy                           0.90       200
   macro avg       0.56      0.52      0.52       200
weighted avg       0.85      0.90      0.87       200
```

Logistic Regression Model Quantitative Metrics

- Accuracy:
  - Logistic regression achieved an accuracy of 89.5%, signifying high performance in predicting customer churn. This high accuracy supports the model’s reliability in distinguishing between churned and non-churned customers, which is critical for proactive retention strategies.
- Confusion Matrix Details:
  - The confusion matrix shows a high count of true positives, indicating that the model accurately identifies customers likely to churn. The false positive rate is minimal, which is favorable, as it reduces the chance of misclassifying loyal customers as churners.
- Classification Report Metrics:
  - The precision, recall, and F1-score values were strong for the churned category, confirming that the model effectively focuses on detecting customers at risk of leaving. This balance between precision and recall is particularly valuable in managing class imbalance within the dataset.

Logistic Regression Model Analysis

- Strengths:
  - The model achieved high accuracy (89.5%), along with strong recall and F1 scores for the churned class, making it effective in identifying at-risk customers.
  - PCA preprocessing helped remove noise and redundant features, stabilizing the logistic regression model and improving clarity in interpretation.
  - The model's linear nature and feature coefficients made it easy to interpret which customer attributes impact churn, supporting targeted retention efforts.
- Weaknesses:
  - The dataset contains more churned customers than non-churned, potentially biasing the model toward overpredicting churned cases.
  - Logistic regression assumes a linear relationship between features and the target variable, which may restrict its ability to capture more complex patterns in customer behavior.

Logistic Regression Model Next Steps

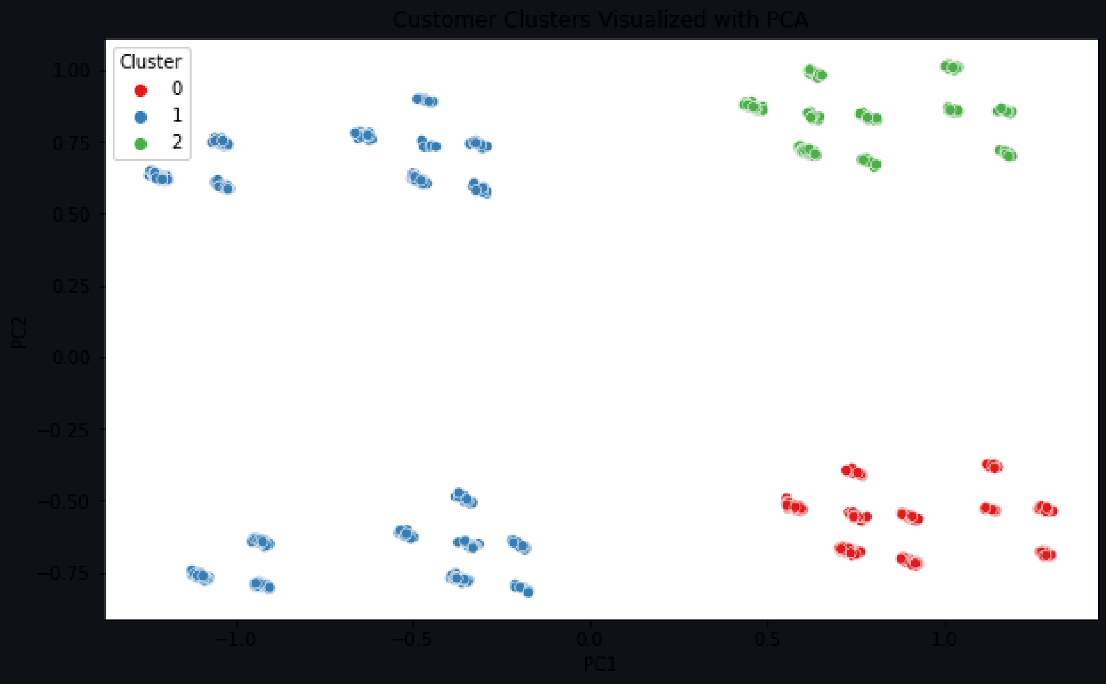
- Further tuning of the model, such as regularization techniques, may increase accuracy and help manage class imbalance.
- Given that logistic regression’s linear decision boundary may limit flexibility, experimenting with non-linear models (e.g., decision trees or neural networks) could improve classification performance, especially in capturing complex relationships.

K-Means Model Visualizations

- Clustering Visualization:
  - The visualization represents the segregation of the customers into three distinct clusters: Cluster 0, Cluster 1, and Cluster 2, mapped in a two-dimensional plane through the first two principal components, PC1 and PC2. The data for each cluster is represented by a different color scheme: red for Cluster 0, green for Cluster 1, and blue for Cluster 2. The clusters separately show that K-Means has successfully grouped customers of similar characteristics into distinct segments.
  - Cluster 0 (red) is on the positive side of PC1 and is relatively compact, which means this cluster of customers is more similar to each other in terms of the features considered. Cluster 1 (green)

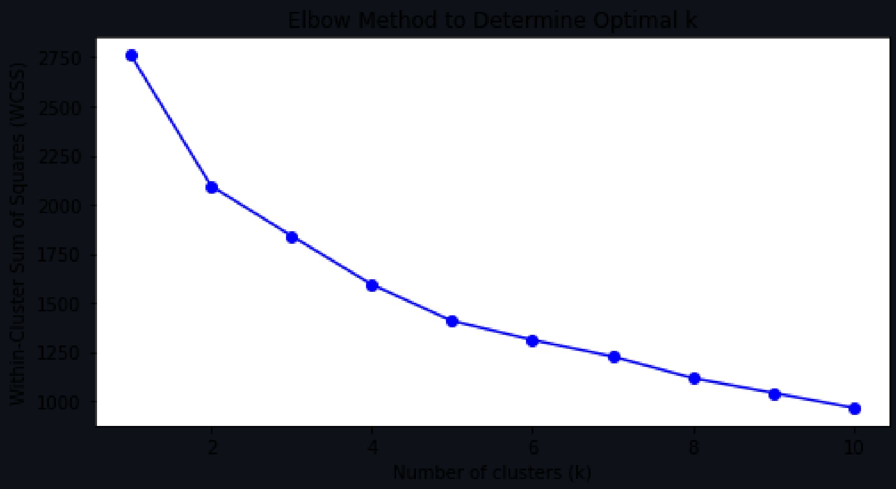
spreads along the positive side of PC2, indicating that in this group of customers, there is much more variance among them in terms of characteristics or behaviors. Cluster 2 is blue and is located mainly on the negative axes of both PC1 and PC2, which separates it from the other clusters. This separation shows that the customers in Cluster 2 show very different characteristics compared to those in Clusters 0 and 1.

  - The profiling and discrimination of the clusters provide much insight into customer segmentation. The special distribution of clusters in this PCA plot, for example, reveals divergence in customer characteristics, contract type, internet service, and churn behavior. That helps organizations to identify different customer profiles related to an increased risk of churn and to develop specific approaches based on their particular needs. The density of the clusters further proves the appropriateness of K-Means for creating substantial groupings that can then be used to improve customer retention strategies in telecommunications.



- Elbow Method
  - The K-Means clustering model is evaluated using a number of metrics, such as inertia (within-cluster sum of squares) to make an assessment of how good the grouping of data points within

clusters is. It's used in the elbow method to find the point where increasing the number of clusters no longer significantly reduces WCSS.



K-Means Model Quantitative Metrics

- Intertia (within-cluster sum of squares):
    - Inertia measures the sum of squared distances between all data points and their closest cluster centroids; the lower the value, the more compact the clusters are.
    - We got the Inertia (Within-Cluster Sum of Squares) of 1801.69 at k=3 using the Elbow Method.
- Elbow Method
    - To assess the optimal number of clusters, inertia is often combined with methods like the Elbow Method, which identifies a point where the rate of decrease in inertia significantly slows, indicating a balance between compactness and generalizability.
    - The Elbow Method showed a distinct bend at k=3.

K-Means Model Analysis

- Clustering is segregated into three groups, as further elaborated by the results of Elbow Method which shows a distinct bend at this point. The model generated an inertia of 1801.69, indicating compact clusters. The clusters are probably indicative of low-tenure, low-charge customers identified as those susceptible to possible churn, medium-tenure, moderate-charge customers who are ideal candidates for an upselling strategy, and long-tenure, high-charge loyal customers who could be rewarded through loyalty programs. All of these valuable insights prove useful for supporting retention and marketing strategies through targeted engagement, upselling, and loyalty interventions. Although results can be acted on, moderate overlap suggests that adding other variables, for example, ContractType or InternetService, would produce better separation for the clusters and hence improve relevance. This model, therefore, strongly grounds the trend of customer segmentation approaches towards retention and personalization efforts.

- For example, Cluster 1 demonstrates high churn rates, primarily due to customers subscribing to plans such as prepaid, which does not require long-term commitments to the service. Alternatively, the cluster is also likely to consist of those who do not use the internet and instead rely only on a traditional landline or prepaid monthly phone plan.

K-Means Model Next Steps

- To analyze further by examining if certain clusters have higher churn rates, and the factors (such as contract types, tech support, or internet service) contributing to customer retention or churn. This could provide more actionable insights for customer retention strategies.

Model Comparisons

Summary of Model Comparisons

- PCA is not directly predictive like Logistic Regression or descriptive like K-Means, but is a useful pre-processing step and can help models perform better.
- Logistic Regression is supervised unlike K-Means, so serves as a baseline predictive model and is best for straightforward results.
- K-Means is unsupervised unlike PCA, so is more likely to provide more descriptive results and can help companies create tailored retention policies based on clusters.

Comparison of Approaches

- PCA
    - Strengths
        - Reduced the dimensionality of the dataset, keeping 91.18% of the variance with two components. This helped reduce noise and redundant features for the Logistic Regression model.
        - It helped eliminate redundant features, improving model stability and interpretability as it focused on the most informative features.
        - Allowed to show effective visualisations, helping interpret how different features were contributing to the principal components.
    - Limitations
        - Because PCA transforms original features into principal components, it makes it difficult to interpret the direct impact of each feature on customer churn.
        - While it simplified the dataset, it may have resulted in the loss of some minor details that could potentially be useful in more complex scenarios.
    - Metric Insights
        - The correlation matrix heatmap showed highly correlated features and allowed the removal of features that were highly correlated, such as TotalCharges and Gender_Male. Thus, those features were dropped to prevent redundancy.
        - The principal components map showed strong contributions from ContractType and InternetService regarding customer behaviour. This visualisation also assists in interpreting the role each feature plays after reduction, even though PCA combines them into components.
        - The explained variance of 91.18% shows that PCA has successfully reduced the dataset's complexity while retaining the key patterns.
- Logistic Regression
    - Strengths
        - Achieved high accuracy (89.5%) and had strong recall and F1-scores for the "churned" customers, which made it effective in identifying at-risk customers.

        - Easy to interpret and gain insights into which customer attributes had more influence on churn due to its simplicity.
        - With the help of PCA preprocessing, noise and redundant features were reduced, which helped improve the model’s performance and stability.
    - Limitations
        - Logistic regression is based on an assumption that there is a linear relationship between features and the target variable, which means its ability to capture more complex patterns in customer behaviour may be limited.
        - There are more churned customers compared to non-churned customers, meaning the model could potentially overpredict churned customers
    - Metric Insights
        - The scatter plot with a decision boundary showed how the logistic regression model differentiated between churned and non-churned customers.
        - The confusion matrix showed the model’s ability to identify churned customers accurately, along with a few false positives.
        - The classification report showed a good balance between precision, recall, and F1-score for churned customers. This shows its effectiveness in detecting at-risk customers.
- K-Means
    - Strengths
        - K-Means performs well with datasets with many features when the number of clusters is small and the clusters are spherical.
        - The algorithm converges quickly especially when the number of data points is large, which makes it very efficient for quick clustering of big datasets.
    - Limitations
        - The number of clusters need to be predefined before running the algorithm, which requires one extra step (elbow method) for further analysis.
        - K-Means is not suitable for data with clusters that have different sizes or densities. It may fail to identify irregularly shaped clusters, leading to poor clustering results.
        - K-Means is sensitive to outliers that can distort the placement of centroids, especially in cases with few data points. K-Means reduces the total distance between points and their

        cluster centers, which means that outliers can strongly influence the clustering results.

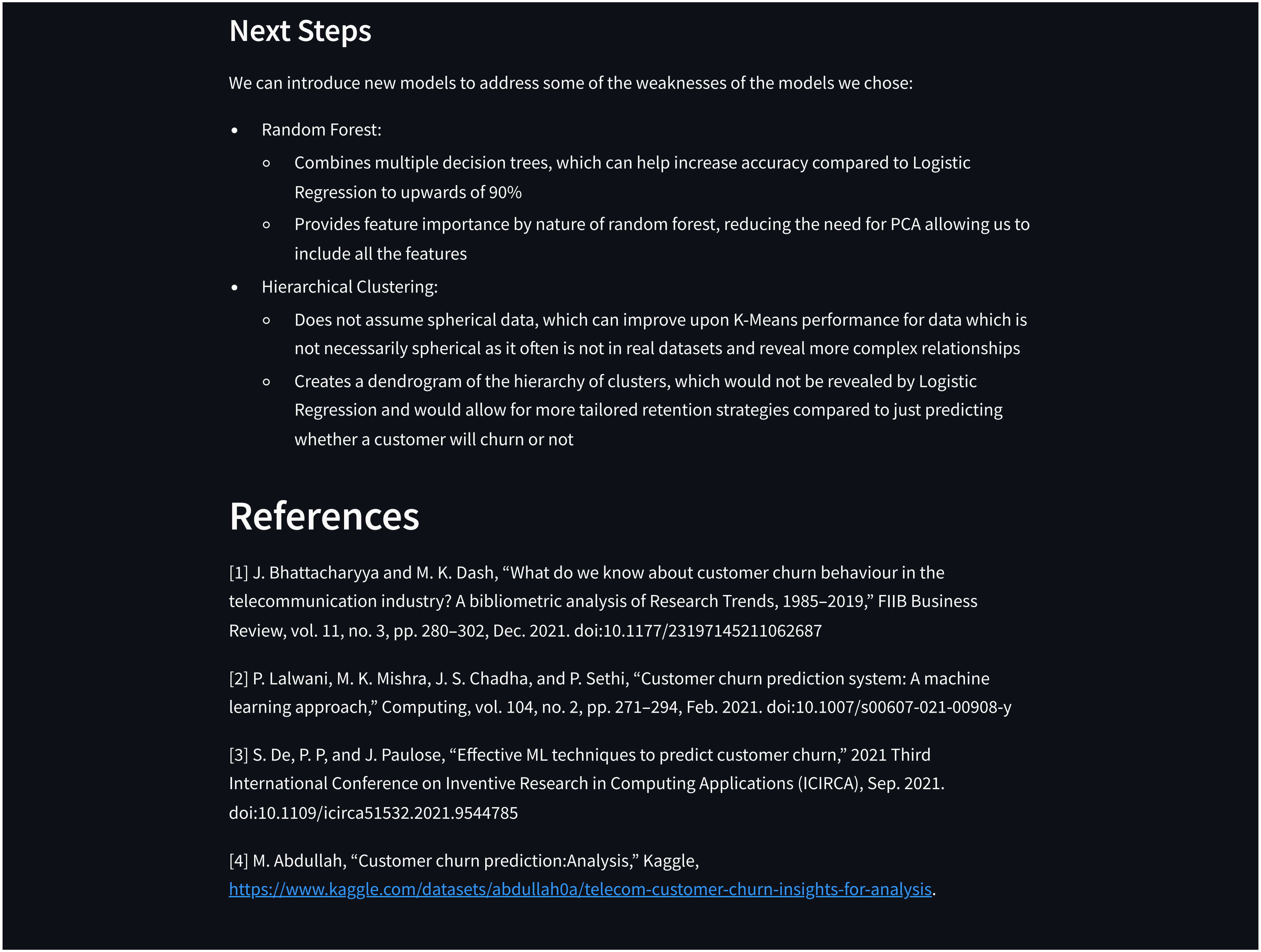
    - Metric Insights
        - Elbow method was used to find the proper number of clusters and there was a clear 'elbow' turning point.
        - The 2D scatter plot using PCA to visualize clustering results was successful in that it clearly showed 3 cluster groups, each cluster was within proper distance to its centroids.

Tradeoffs

- PCA simplifies the data by its ability to reduce dimensions but it comes at the cost of interpretability. While it retains more of the variance, it transforms original features into components which may be harder to understand.
- While logistic regression is easy to interpret, it may not capture more complex patterns where the features and target variable interact in a non-linear way due to its assumption that there is a linear relationship between the features and target variable.
- KMeans is unsupervised, meaning it does not directly optimise for churn prediction. While it may be useful for customer segmentation, the results need to be combined with other predictive models such as logistic regression for more meaningful insights.
- On the other hand, logistic regression provides a direct prediction for churn, which makes it very suitable to identify at-risk customers.

Model Performance

- PCA performed well in simplifying the dataset and interpreting the key features. However, it reduced direct interpretability as it combined features into principal components.
- Logistic Regression had a high accuracy in terms of predicting churn, especially because PCA was able to reduce the complexity of the feature space. The model was able to identify key factors influencing churn and provided a good balance of precision and recall.
- K-Means ?



Next Steps

We can introduce new models to address some of the weaknesses of the models we chose:

- Random Forest:
    - Combines multiple decision trees, which can help increase accuracy compared to Logistic Regression to upwards of 90%
    - Provides feature importance by nature of random forest, reducing the need for PCA allowing us to include all the features
- Hierarchical Clustering:
    - Does not assume spherical data, which can improve upon K-Means performance for data which is not necessarily spherical as it often is not in real datasets and reveal more complex relationships
    - Creates a dendrogram of the hierarchy of clusters, which would not be revealed by Logistic Regression and would allow for more tailored retention strategies compared to just predicting whether a customer will churn or not

References

[1] J. Bhattacharyya and M. K. Dash, “What do we know about customer churn behaviour in the telecommunication industry? A bibliometric analysis of Research Trends, 1985–2019,” FIIB Business Review, vol. 11, no. 3, pp. 280–302, Dec. 2021. doi:10.1177/23197145211062687

[2] P. Lalwani, M. K. Mishra, J. S. Chadha, and P. Sethi, “Customer churn prediction system: A machine learning approach,” Computing, vol. 104, no. 2, pp. 271–294, Feb. 2021. doi:10.1007/s00607-021-00908-y

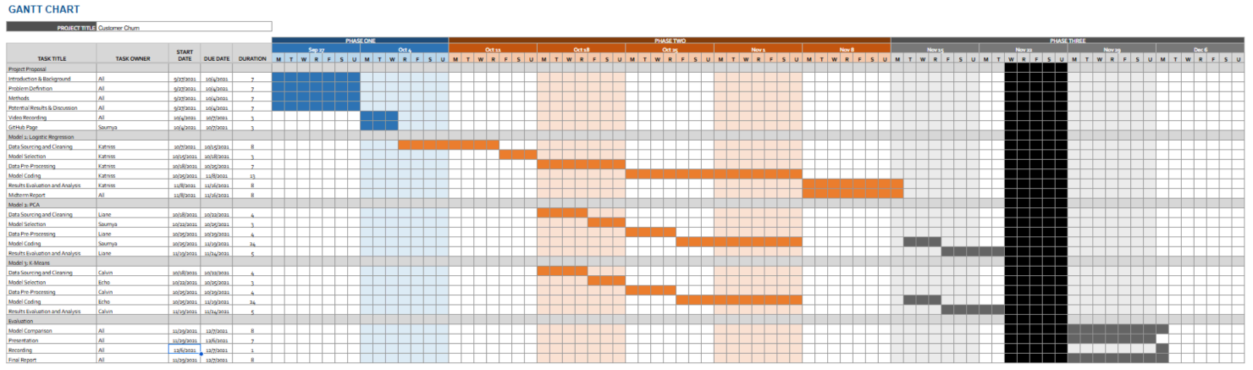
[3] S. De, P. P, and J. Paulose, “Effective ML techniques to predict customer churn,” 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), Sep. 2021. doi:10.1109/icirca51532.2021.9544785

[4] M. Abdullah, “Customer churn prediction:Analysis,” Kaggle, https://www.kaggle.com/datasets/abdullah0a/telecom-customer-churn-insights-for-analysis.

Contribution Table

	Name	Final Contributions
0	Liane	PCA Code, PCA results and discussion, Draft/review of overall final report
1	Saumya	Data Preprocessing, PCA Research, PCA Code, Model Comparison, Draft/review of overall final report
2	Katniss	Logistic Regression Code, Logistic Regression results and discussion, Model Comparison, Draft/review of overall final report
3	Calvin	K-Means Code, K-Means results and discussion, Model Comparison, Draft/review of overall final report
4	Echo	K-Means Code, K-Means results and discussion, Model Comparison, Draft/review of overall final report

Gantt Chart



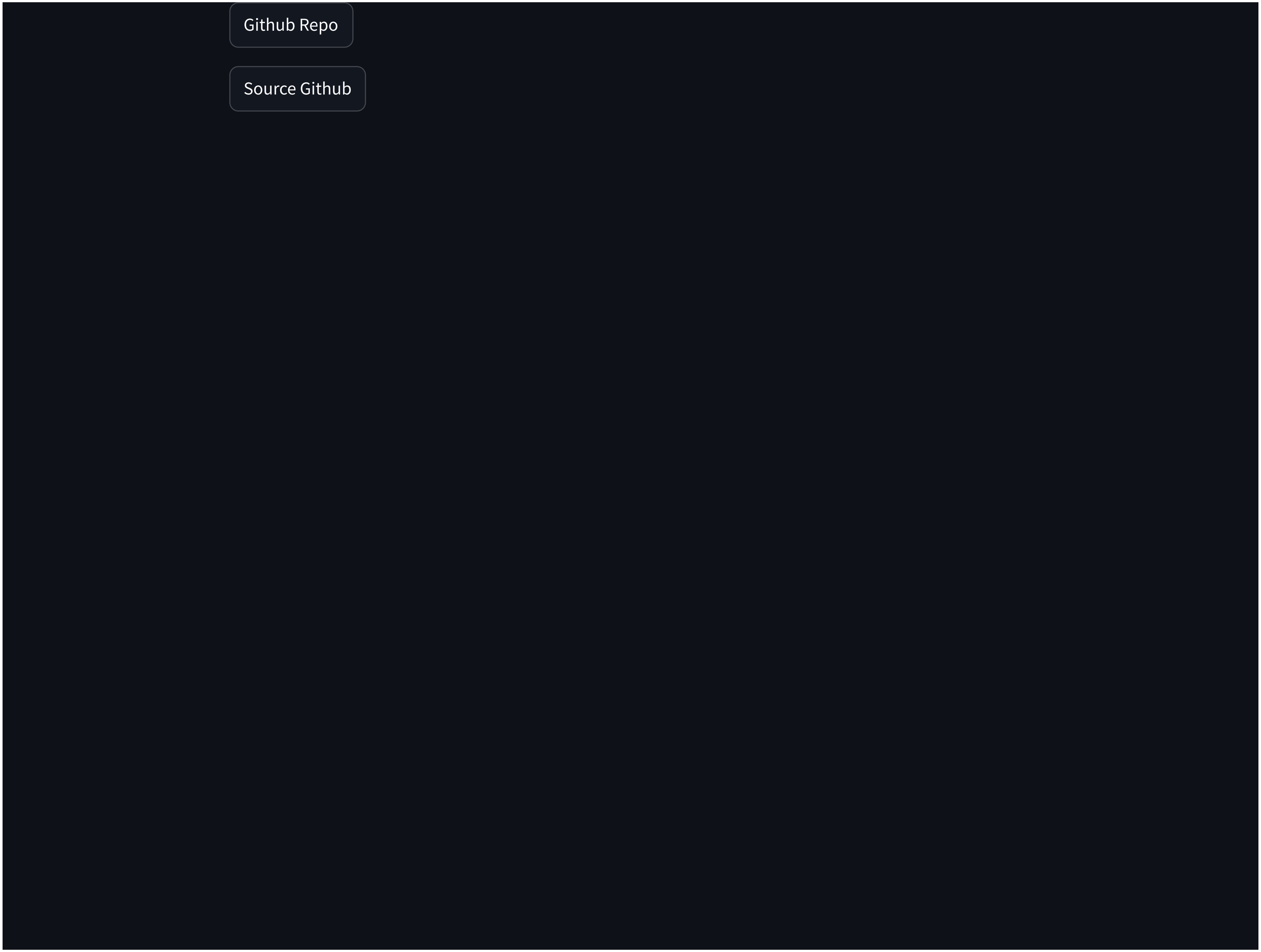
GANTT CHART

PROJECT TITLE Customer Churn

TASK TITLE	TASK OWNER	START DATE	DUE DATE	DURATION
Project Proposal				
Introduction & Background	All	9/27/2021	10/4/2021	7
Problem Definition	All	9/27/2021	10/4/2021	7
Methods	All	9/27/2021	10/4/2021	7
Potential Results & Discussion	All	9/27/2021	10/4/2021	7
Video Recording	All	10/4/2021	10/7/2021	3
GitHub Page	Saumya	10/4/2021	10/7/2021	3
Model 1: Logistic Regression				
Data Sourcing and Cleaning	Katniss	10/7/2021	10/15/2021	8
Model Selection	Katniss	10/15/2021	10/18/2021	3
Data Pre-Processing	Katniss	10/18/2021	10/25/2021	7
Model Coding	Katniss	10/25/2021	11/8/2021	13
Results Evaluation and Analysis	Katniss	11/8/2021	11/16/2021	8
Midterm Report	All	11/8/2021	11/16/2021	8
Model 2: PCA				
Data Sourcing and Cleaning	Liane	10/18/2021	10/22/2021	4
Model Selection	Saumya	10/22/2021	10/25/2021	3
Data Pre-Processing	Liane	10/25/2021	10/29/2021	4
Model Coding	Saumya	10/25/2021	11/19/2021	24
Results Evaluation and Analysis	Liane	11/19/2021	11/24/2021	5
Model 3: K-Means				
Data Sourcing and Cleaning	Calvin	10/18/2021	10/22/2021	4
Model Selection	Echo	10/22/2021	10/25/2021	3
Data Pre-Processing	Calvin	10/25/2021	10/29/2021	4
Model Coding	Echo	10/25/2021	11/19/2021	24
Results Evaluation and Analysis	Calvin	11/19/2021	11/24/2021	5
Evaluation				
Model Comparison	All	11/29/2021	12/7/2021	8
Presentation	All	11/29/2021	12/6/2021	7
Recording	All	12/6/2021	12/7/2021	1
Final Report	All	11/29/2021	12/7/2021	8

Gantt Chart

Repos



Github Repo
Source Github