Customer Churn Prediction

Introduction/Background

Our topic is predicting customer churn, which is the percentage of customers who stop using a company's products or services over a set period of time. Predicting customer churn can be vital for a business to understand why customers leave. Churn prediction has become a critical focus within the machine learning community, with numerous studies exploring various techniques to maximize predictive accuracy. Lalwani et al. applied ensemble methods like AdaBoost and XGBoost, showing these models' strengths in identifying nonlinear patterns in telecom data, while Sharma and Panigrahi leveraged Artificial Neural Networks to achieve over 92% accuracy, capitalizing on the complexity these networks can capture in high-dimensional datasets. Karahoca and Karahoca took a different approach, using unsupervised clustering to identify churn-prone customer segments with 93% correctness. Ongoing challenges identified in current literature include data quality and imbalance, generalization across different industries, and optimizing data gathering for feature selection.

The importance of this issue lends itself to the availability of a lot of high-quality data. For our project, we will use the Customer Churn Prediction: Analysis dataset from Kaggle (https://www.kaggle.com/datasets/abdullah0a/telecom-customer-churn-insights-for-analysis), which provides data on over 1,000 customers in the telecom industry, including key features like customer age, gender, tenure, and monthly charges.

Problem Definition

Customer churn is the percentage of customers who stop using a company's products or services over a set period of time. It is a key metric used to analyze companies' health, especially in subscription-based models, which many companies have recently shifted to to increase revenue streams. In these subscription-based models, high customer churn means a direct loss of revenue, negatively impacting a company’s bottom line. Beyond financial impact, churn often signals deeper issues like customer dissatisfaction, which can lead to an eroding brand reputation and can potentially deter new prospective customers. Additionally, high churn rates can undermine growth efforts, as acquiring new customers is often more costly than retaining old ones. This all makes customer churn a vital metric necessary to evaluate a company’s overall health, growth, and sustainability. Therefore, it is crucial for businesses to be able to evaluate and predict customer churn ahead of time before risking customers actually leaving, and to prevent it by understanding exactly why customers are leaving.

Methods

Preprocessing Steps

To make sure our models are effective, we plan to preprocess the data as such:

- Handling Missing Data: We will fill/replace or remove missing values in the dataset.

- Feature Encoding: Categorical features like gender and contract type will be converted to numerical representations of them through one-hot encoding.
- Standardization: Continuous features such as monthly charges and tenure will be standardized to have unit variance and zero mean across all features, which is important for PCA (for preprocessing), as well as models like SVM and Logistic Regression.
- PCA: We use PCA to reduce the dimensionality of the data down to the most important principal components (capturing 95% of the variance from the original data). This is to ensure that we reduce model complexity, preserve relevant data, and de-noise our data without removing too much information.

Machine Learning Models

We will implement the following ML models to predict customer churn:

- **Logistic Regression:** Will serve as our baseline algorithm for binary classification (relatively effective). Its primary benefit is interpretability, helping us understand how each feature (e.g. tenure, monthly charges) impacts churn probability.
- **Random Forest:** Good at improving prediction accuracy by aggregating multiple decision trees, which are more flexible/intuitive compared to logistic regression. This model performs well on both categorical (e.g. contract type, gender) and continuous features (e.g. monthly charges), and is well equipped to handle missing/noisy data. Allows us to capture complex decision boundaries and interactions between features, ideal for non-linear relationships.
- **Support Vector Machine:** Effective for binary classification tasks with a distinct margin between the classes, this fits well with our use case of a customer either churning or not churning. Strong at dealing with smaller datasets and maximizing separation of groupings which we need for our dataset.
- **K-Means Clustering:** Unsupervised model to group data in natural clusters. It's useful to better understand our dataset by analyzing patterns in the data and grouping customers with similar characteristics. These groupings of customers can help analyze the characteristics of what types of customers are more/less likely to churn

Results and Discussion

We will evaluate our models using the following metrics:

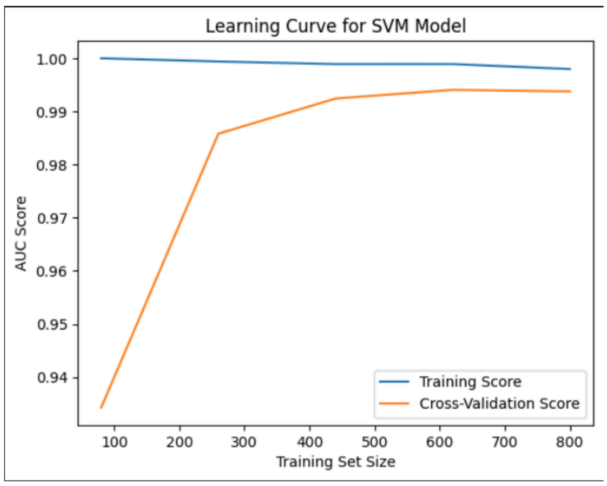
- Accuracy: Measure correct model prediction; can be misleading if the dataset is imbalanced.
- F1 Score: Balanced metric combining precision and recall, minimizing false positives/negatives—critical in churn prediction where both types of errors have business consequences.
- ROC Curve & AUC: To assess how well the model distinguishes between churners and non-churners across different thresholds. The area under ROC curve (AUC) will help evaluate the model's ability on how well it classifies data points into categories.

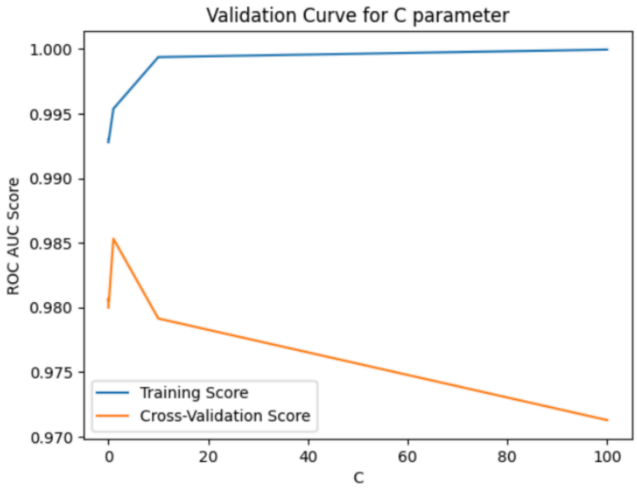
We expect Random Forest and SVM to outperform Logistic Regression when comparing their precision and recall in detecting churn rate. Our goal is to achieve at least 92-93% accuracy, something similar to the results we have seen in previous studies.

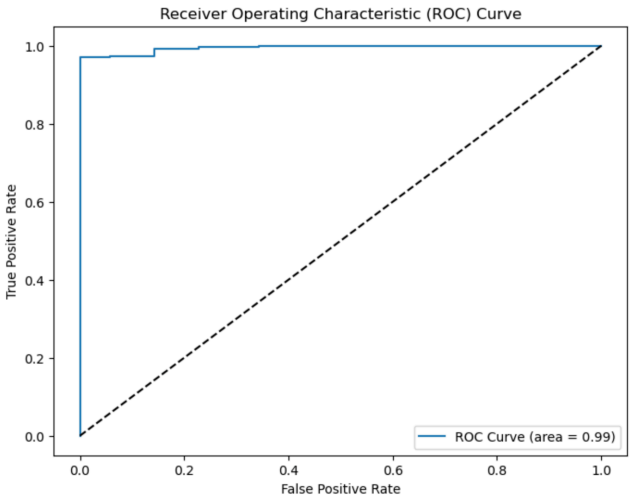
We will also test the impact of unsupervised ML techniques like K-Means clustering to identify potential customer groups as churners/non-churners. However, these results are secondary and more supplementary to the main results above.

SVM Results

Visualizations







Confusion Matrix on Test Set:

	Positive (predicted)	Negative (predicted)
Positive (actual)	30	5
Negative (actual)	2	263

Quantitative Results

- Accuracy: 0.98
- F1 Score: 0.987
- Cohen's Kappa Score: 0.88
- Specificity: 0.86
- ROC-AUC: 0.99

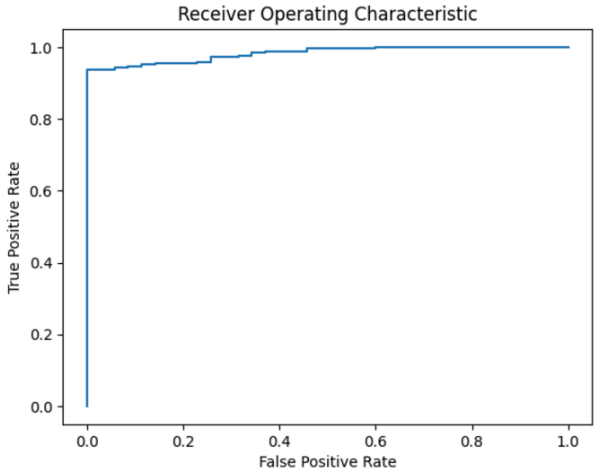
Analysis

SVM performed extremely well, with the ability to predict churn for a given customer with 0.98 accuracy. After tuning the hyperparameters, we found a value of 10 for C yielded the most optimal results. An accuracy of 98% means our model's predictions are correct and is making few mistakes. However, we recognize that the data is unbalanced, since there are significantly more data points for customers who have churned compared to customers who haven't. Leveraging the F1 score provides more robustness over accuracy, providing a measure of how the model classifies positive cases and avoids false positives or negatives. Our model achieved an F1 score of 0.987, indicating a good balance between precision and recall. Cohen's Kappa Score, which is from -1 to 1, also shows our model had strong agreements between

the model's predictions and the true labels. The Specificity of 0.86 shows that 86% of non churn cases are correctly identified. An ROC-AUC of 0.99 shows the model separates classes across various thresholds - since the maximum possible score is 1, this shows excellent separability.

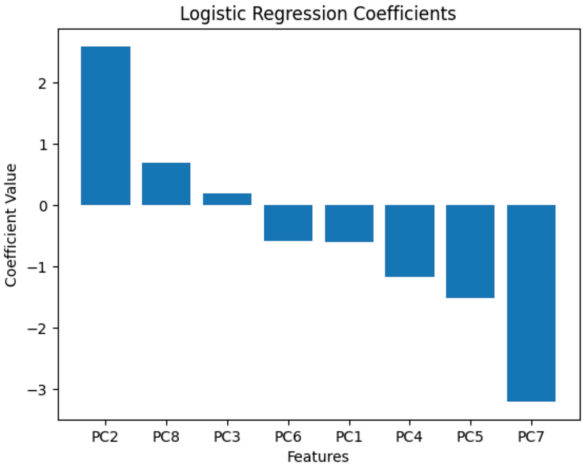
Logistic Regression Results

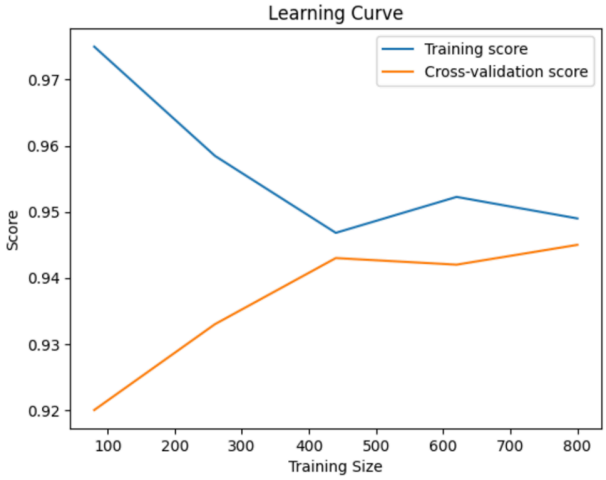
Visualizations



Confusion Matrix on Test Set:

	Positive (predicted)	Negative (predicted)
Positive (actual)	24	11
Negative (actual)	7	258





Quantitative Results

```
Classification Report:
              precision    recall  f1-score   support

           0       0.77      0.69      0.73        35
           1       0.96      0.97      0.97       265

    accuracy                           0.94       300
   macro avg       0.87      0.83      0.85       300
weighted avg       0.94      0.94      0.94       300
```

- Accuracy: 0.94

- F1 Score: 0.85
- ROC-AUC: 0.98

Analysis

Logistic regression also performed well but not quite as well as SVM did, as expected, ending with a 94% accuracy. However, since our dataset is imbalanced with more customers leaving (churning) then not, F1 score is a more accurate metric as it factors in precision and recall, accounting for data imbalance. In this case the F1 score is 0.85 which is considerably lower than in SVM. Taking a closer look at the results, it's clear that the model has a high precision and recall for the customers that do churn but a far lower precision and recall for customers who don't churn. This is likely because of the imbalance of data in our dataset. SVM however still had a high precision and recall for both the customer churn classes and non customer churn classes despite the lack of data. This is likely due to how SVM works to separate the data via a hyperplane that maximizes the margin, which is more effective for smaller datasets like in our case where the non customer churn case isn't highly represented in the dataset. The arguably most interesting and useful information gathered from the logistic regression is from the coefficients developed from training the model, which can be seen in the graph above, since we can use this information to determine which features in the dataset were the most impactful contributors to the result of churn or no churn. Since the regression was run on the dataset after Principal Component Analysis (PCA) was performed, the coefficients correspond to the PCA components rather than features directly, ordered by how much variance in the dataset can be explained by each (PC1 meaning most variance can be explain, PC8 meaning least variance can be explained):

```
Top contributing features to PC1: ['TotalCharges', 'Tenure', 'MonthlyCharges']
Top contributing features to PC2: ['TechSupport', 'InternetService_Fiber Optic', 'Age']
Top contributing features to PC3: ['ContractType_Two-Year', 'ContractType_One-Year', 'Gender']
Top contributing features to PC4: ['Gender', 'Age', 'MonthlyCharges']
Top contributing features to PC5: ['MonthlyCharges', 'Age', 'Tenure']
Top contributing features to PC6: ['Gender', 'Age', 'InternetService_Fiber Optic']
Top contributing features to PC7: ['ContractType_One-Year', 'ContractType_Two-Year', 'InternetService_Fiber Optic']
Top contributing features to PC8: ['TechSupport', 'InternetService_Fiber Optic', 'ContractType_Two-Year']
```

Based on the graph, the logistic regression coefficients that have the largest magnitude are PC2 and PC7, which have a large positive and large negative value respectively. This means that PC2 is a large factor in a yes for customer churn (a customer leaving), while PC7 is a large factor in a no for customer churn (a customer staying). Based on the images above the top contributing features to PC2 are tech support (yes/no whether or not the customer has tech support), internet service fiber optic (yes/no whether or not the customer has fiber optic vs. DSL - broadband internet over telephone lines), and age (age of the customer). This image below shows how each of these features contribute to PC2:

```
Correlations of top contributing features to PC2:
  TechSupport: -0.3620
  InternetService_Fiber Optic: -0.3581
  Age: 0.1219
```

This makes sense, as a yes value for tech support means that the customer is less likely to churn, since they will receive more help from the company. This is because tech support has a negative weight for PC2, meaning the larger it is the smaller PC2 is, and the smaller PC2 has a large impact on the customer not churning. Similarly for internet service fiber optic, a yes value in this category means the customer is less likely to churn which also makes sense as fiber optic offers much faster internet than the alternative for this category, DSL. From these two features, it seems that customers who have the best services from the company (tech support and the fastest internet, fiber optic) are less likely to leave. Therefore in order to prevent customer churn, companies should seek to enlist customers in their best and most rewarding services. Additionally, age has a positive correlation with PC2 meaning that a larger age means a customer is more likely to churn and leave. This could be for a variety of reasons, such as the customer retiring and no longer needing fast internet. Based on this feature, it seems that younger customers are more likely to remain with a company rather than leave. This leads to the conclusion that companies should seek a younger customer base to reduce churn.

The top contributing features to PC7 based on the image above are one year contracts (yes/no if the customer has a one year contract), two year contracts (yes/no if the customer has a two year contract), and internet service fiber optic. This image below shows how each of these features contribute to PC7:

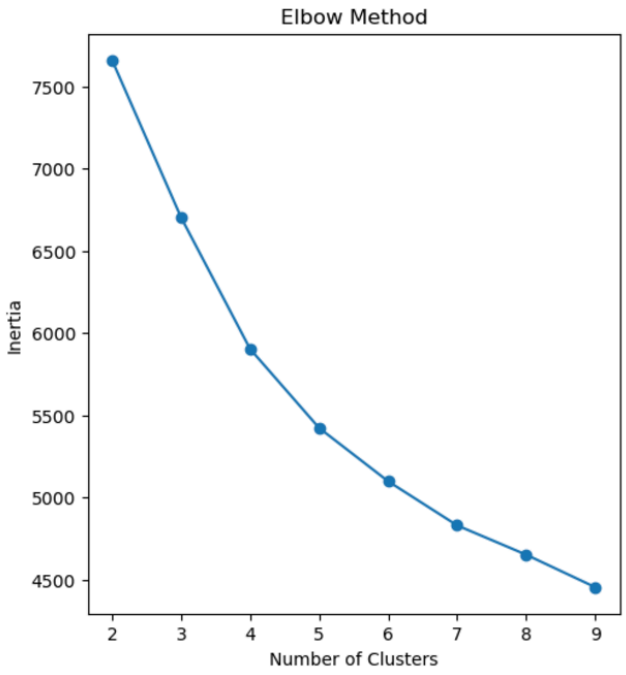
```
Correlations of top contributing features to PC7:
  ContractType_One-Year: 0.3289
  ContractType_Two-Year: 0.3284
  InternetService_Fiber Optic: -0.1360
```

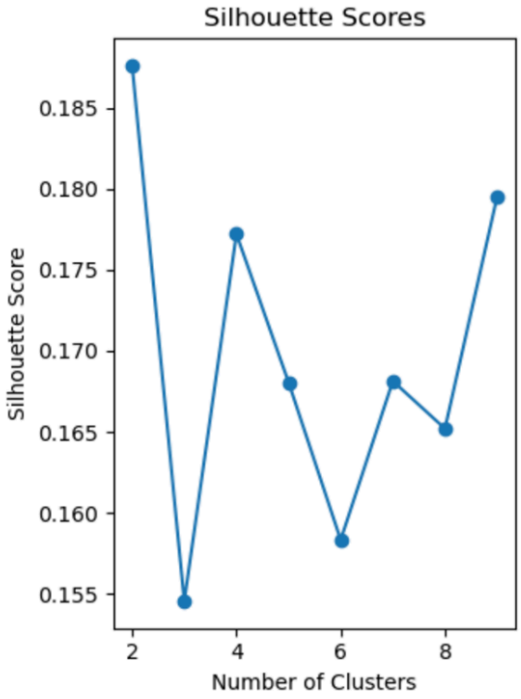
The other alternative for the contract type feature of the dataset is a month to month contract rather than a one year or two year one. Therefore, these results make sense as if a customer is locked in to a one year or two year contract, they're far less likely to churn or leave as if they do they'd be breaking the contract and presumably have to pay out the rest of it or face legal consequences. This reasonably leads to the conclusion that another effective way to prevent customer churn is to keep customers engaged in long term contracts rather than having them pay monthly subscriptions which they can cancel at any point.

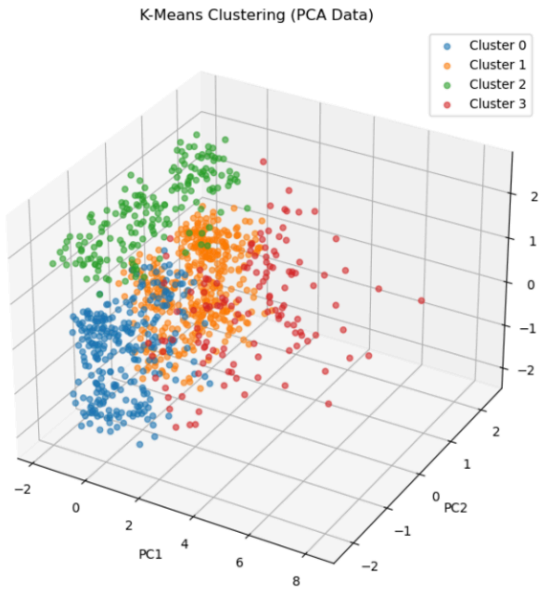
In conclusion, it's safe to say that the reasons that contribute to a customer leaving or not are highly dependent on the business the company is in. However, based on this analysis, we can conclude that keeping customers happy with high-end services and locking customers into long term deals rather than a monthly subscription are safe ways to reduce customer churn and keep people coming back to a business.

K-Means Clustering Results

Visualizations







(vertical axis is PC3)

Quantitative Results

```
Cluster Distribution:
Cluster
1    408
0    278
2    182
3    132
Name: count, dtype: int64
```

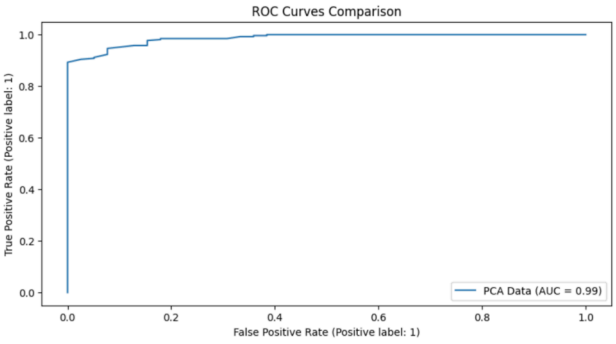
```
Churn Percentage by Cluster:
Churn             0          1
Cluster
0         13.669065  86.330935
1          5.392157  94.607843
2         21.428571  78.571429
3         13.636364  86.363636
```

Analysis

K-means clustering was used to attempt to group the data into natural clusters and potentially help organize customer data into overall categories to help improve the understanding of our dataset. Additionally we had hoped that the clusters may form such that each cluster had a different majority of no churn or churn to help understand what types of customers fit into churn or no churn. However, as seen by the quantitative metrics above, this was not the case based on the results of the clustering using k-means. This is particularly seen in the churn percentage by cluster metric, which displays how the overwhelming majority of every cluster is churn. This is likely due to the fact that the majority of the data in our dataset is yes churn due to the imbalance of class representation. A value of 4 was chosen for the number of clusters based on the elbow method and silhouette score. Although the elbow method didn't produce a sharp edge to select the number of clusters with, 4 clusters was where the rate of inertia decrease seemed to slow down the most. Silhouette score is used to measure the quality of a clustering by evaluating how similar a data point is to its own cluster, and a higher score is better. Although 2 clusters had the highest silhouette score, 4 clusters were chosen since 4 clusters still had a relatively high silhouette score and was the best option based on the elbow method as well. The graphs for the clustering do show distinct groups, however it's difficult to garner if there are specific features correlated with each cluster distinctly enough to categorize the customers in the dataset in the groups. Since the majority of each cluster is of the yes churn class, this analysis won't provide any real insight into solving the problem of limiting customer churn. Further analysis needs to be done potentially leveraging different clustering methods such as DBSCAN that may produce better results and more distinct clusters.

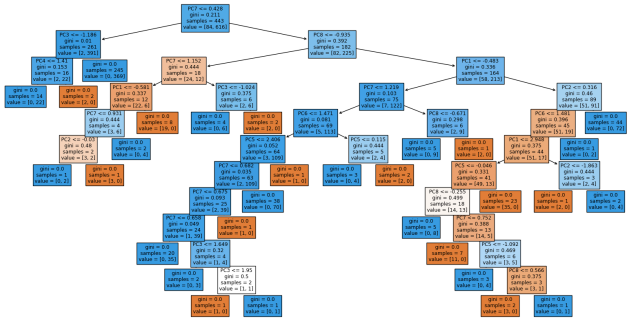
Random Forest

Visualizations



Confusion Matrix on Test Set:

	Positive (predicted)	Negative (predicted)
Positive (actual)	37	2
Negative (actual)	6	255



Quantitative Results

```
Classification Report (Resampled):
              precision    recall  f1-score   support

           0       0.86      0.95      0.90        39
           1       0.99      0.98      0.98       261

    accuracy                           0.97       300
   macro avg       0.93      0.96      0.94       300
weighted avg       0.98      0.97      0.97       300
```

- Accuracy: 0.97
- F1 Score: 0.94
- ROC-AUC: 0.99

Analysis

Random forest performed very well with a 0.97 accuracy, nearly as high as SVM. Additionally, its F1 score is excellent as well for both precision and recall indicating that the model performed well for both classes of churn and no churn despite the lack of no churn data. However, it wasn't quite as good as SVM which had an F1 score of 0.98 rather than random forest at 0.94. The ROC-AUC score was excellent and the same as SVM at 0.99 meaning there was good separability between the classes for the classification. It makes sense that the random forest model performed well since it works well for accounting for non linear decision boundaries and factoring in complex relationships in the data that SVM was able to account for through transformations of the data but logistic regression wasn't able to account for as strongly. Additionally, these results were gathered after using SMOTE on the data to help account for the underrepresented no churn class. SMOTE works to generate synthetic samples of the unrepresented data to balance out the dataset. Before running SMOTE, random forest produced the results below:

```
              precision    recall  f1-score   support

           0       0.93      0.67      0.78        39
           1       0.95      0.99      0.97       261

    accuracy                           0.95       300
   macro avg       0.94      0.83      0.87       300
weighted avg       0.95      0.95      0.95       300
```

The F1 score is noticeably lower (macro avg), which makes sense since there was less data for the no churn class. Using SMOTE resulted in a noticeable improvement in the results of the model, indicating it may be useful in other methods we've used as well that struggled due to the class imbalance such as logistic regression. Random forest generally aligns well with our dataset since it's more robust to imbalanced datasets due to its ensemble nature which allows different trees in the forest to focus on different parts of the data, however SMOTE still saw a noticeable improvement.

Comparison of Models and Conclusions

In conclusion, the models that were more robust to imbalanced datasets performed the best, those being SVM and Random Forest, while logistic regression struggled due to the imbalanced data. Running SMOTE for the Random Forest model resulted in better results due to a more balanced dataset, and may be a useful technique to apply to other models that perform poorly on imbalanced datasets like logistic regression. While SVM and Random Forest proved to be the best for classification of new data, logistic regression along with PCA provided valuable insight into what features of our data had the largest effect on whether a customer left/churned or not. Based on the analysis of the weights with the largest

magnitude from the regression and the features that were weighted most in each principal component, it was concluded that customer churn is very business dependent, and companies should make their own models on their own customer data to get the most accurate insights. However in general it was found that customers with the best high-end services from the company were more likely to stay, in addition to customers that were locked in to long term contracts rather than paying monthly subscriptions which can be canceled at any point. Additionally, younger customers were more likely to remain and not churn.

Next Steps

Looking forward, further analysis into customer churn on datasets from different industries could prove as insightful, as this analysis was just on data from the telecom industry. It could be that in different industries, different metrics are more important to customer churn than the metrics found to be most important in this analysis. Additionally, finding a larger dataset with more balanced classes could prove valuable in training more robust models that can more accurately classify new input data. However, running SMOTE on the imbalanced dataset used in this analysis to synthetically generate more data for the imbalanced class did prove to have a positive impact on the model's results, and could be a technique further leveraged in future analysis for models that don't traditionally perform well on imbalanced data. Regarding the clustering using k-means, other clustering algorithms can be used to potentially construct more distinct groups based on features of the dataset, to determine if there are specific archetypes of customers that are more likely to churn or not churn based on the features of the cluster they belong too since k-means produced poor results in this area. These algorithms could include DBSCAN or hierarchical clustering.

References

- Lalwani, Praveen, et al. "Customer Churn Prediction System: A Machine Learning Approach." Computing, vol. 104, no. 2, Feb. 2022, pp. 271–94. Springer Link, https://doi.org/10.1007/s00607-021-00908-y.
- Sharma, Anuj, and Prabin Kumar Panigrahi. "A Neural Network Based Approach for Predicting Customer Churn in Cellular Network Services." International Journal of Computer Applications, vol. 27, no. 11, Aug. 2011, pp. 26–31. DOI.org (Crossref), https://doi.org/10.5120/3344-4605.
- Karahoca, Adem, and Dilek Karahoca. "GSM Churn Management by Using Fuzzy C-Means Clustering and Adaptive Neuro Fuzzy Inference System." Expert Systems with Applications, vol. 38, no. 3, Mar. 2011, pp. 1814–22. DOI.org (Crossref), https://doi.org/10.1016/j.eswa.2010.07.110.

Gantt Chart Link: https://1drv.ms/x/s!As53W-C4_NHTrjI6qPpe4Ggp2Dv5?e=tFunJX

Contributions

Name	Final Contributions
Nathan Gong	Logistic Regression

Ronak Agarwal	SVM
Shyamanth (Sunny) Kudum	Random Forest & KMeans
Alexander Steele	Results Analysis & Presentation
Arnav Patidar	Model final touches, results analysis & presentation