**Predicting the Customer Churn for Telco**

**Abstract**

The goal of this project was to use a classification model to predict customer churn for Telco, a fictitious telecommunications company, and to identify areas within Telco business that could be improved to help retain customers. The basis of the analysis was churn data acquired from [Kaggle](https://www.kaggle.com/datasets/ylchang/telco-customer-churn-1113).

**Business Need**

Telco is a fictitious telecom business that provides telephone and internet services. Churn is the measure of how many customers stop using a product or service. In general, it costs Telco about five times more to attain a new customer as it does to keep an existing customer. Minimizing churn rate will help Telco increase its revenue.

**Data**

The dataset contains 7,024 customer records with 33 features for each customer. Twenty-nine of the features are categorical. In initial data cleaning, eight of the features were determined to be duplicative and dropped, and 25 features were used to inform baseline models and feature engineering.

**Algorithms**

*Feature Engineering*

1. Extensive feature engineering was completed in ArcGIS to derive a new feature representing churn hotspots or the number of churned customers within one mile.
2. Converted categorical features to binary dummy variables using get dummies. For contract type, I manually dropped a class after using get dummies because I was interested in looking at month-to-month contracts as a variable.
3. Removed highly correlated variables. For this iteration, I removed “Total charges” which was highly correlated to “Monthly charges” as well as “Tenure (months)” variables

*Models: Evaluation and Selection*

Logistic regression, decision tree, random forest, and naïve bayes classifiers were fitted and compared based on their F1 scores. F1 score was chosen as the overall evaluation metric because it is the harmonic mean of precision (% of churns predicted that actually were churns) and recall ( % of actual churns).

Random Forest had the strongest cross-validation performance after model tuning and correcting for sample imbalance. Random forest feature importance ranking was used directly to inform business for Telco that could be improved to help retain customers

**Final random forest 10-fold CV scores:**  24 features (4 numeric, 24 categorical)

* F1 - 84%
* Precision - 86%
* Recall – 82%

**Most important features**

1. Number churned customers within mile
2. Tenure (months)
3. Contract (month to month)
4. Monthly charges
5. Customer lifetime value

**Business Recommendations**

1. The cost of products and services is more common in areas with higher numbers of churned customers. Telco should look at minimizing installation and servicing of fiber optic lines, one of its priciest products, to increase the return on investment that can be passed onto the customer
2. Most churned customers have month-to-month contracts. Telco should offer incentive packages to encourage customers to sign annual or two-year contracts. For example, the monthly charge for customers who sign an annual contract could be reduced by $10 per month and $15 per month for two-year contracts.
3. Monthly charges are higher for churned customers with phone services. Telco should collect more information, specifically on their overage charges and information on city taxes and fees, for their services on this subset of customers to determine why they are paying more for standard services.

**Tools**

* Numpy and Pandas for exploratory data analysis and feature engineering
* Scikit-learn for modeling
* Matplotlib and Seaborn for data visualization
* ArcGIS for feature engineering