Explanatory Note for Project:

Project Detail and Description:

**Context**

"Predict behaviour to retain customers. You can analyse all relevant customer data and develop focused customer retention programs." [IBM Sample Data Sets]

**Content**

Each row represents a customer, each column contains customer’s attributes described on the column Metadata.

**The data set includes information about:**

* Customers who left within the last month – the column is called Churn
* Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
* Customer account information – how long they’ve been a customer, contract, payment method, paperless billing, monthly charges, and total charges
* Demographic info about customers – gender, age range, and if they have partners and dependents

From: <https://www.kaggle.com/blastchar/telco-customer-churn>

Reasoning:

As mentioned in the readme, I chose this dataset because I was particularly interested in its characteristics and it’s application to my field, actuarial studies. Both insurance and telephone contracts tend to be long-term in nature, and ideally want to keep their customers for long periods of time, and are particularly interested in the factors that drive them away from the service.

As a result, I thought it would be an interesting foray into the machine learning sphere using python. I ultimately settled with Gradient Boosted Decision Trees as a model as it showed the best performance of the ensemble methods I chose to use for the dataset, and I felt that the differences between the base models for AdaBoost and XGBoost were not that different to explore hyper-parameter tuning and performance testing for those particular models.