WILL THE CUSTOMER CHURN?

- Final project for Ds 105

SCENERIO

As one of the leading telco in the country, the company are proud to provide quality services that have been receiving satisfactory feedback from customers frequently. But from recent quarterly reports, there have been a drop in the customer base.

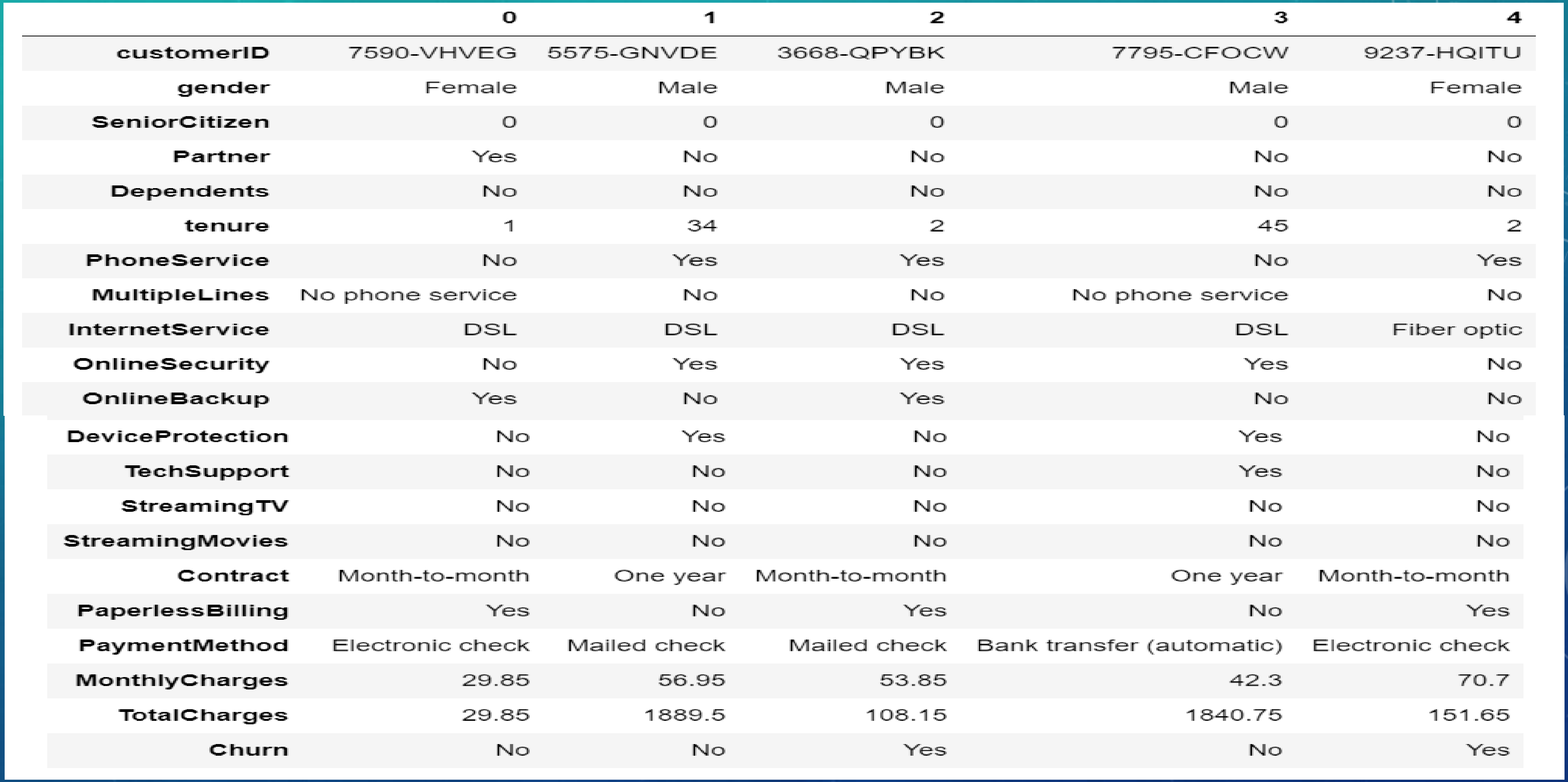
We are instructed by the management to create a model to predict if a customer will churn so that they are able to come up with solutions to keep hold of them.

OBJECTIVES

➢To create a model to predict if the customer will stay or leave(churn) by:

- Understanding the data
- Analysis the features to see the relationship with the label
- Pre-process the data for the model
- Find the baseline of some models and choose the suitable model
- Evaluate the final model selected

DATASETS (TELCO CUSTOMER CHURN | KAGGLE)



	0	1	2	3	4
customerID	7590-VHVEG	5575-GNVDE	3668-QPYBK	7795-CFOCW	9237-HQITU
gender	Female	Male	Male	Male	Female
SeniorCitizen	0	0	0	0	0
Partner	Yes	No	No	No	No
Dependents	No	No	No	No	No
tenure	1	34	2	45	2
PhoneService	No	Yes	Yes	No	Yes
MultipleLines	No phone service	No	No	No phone service	No
InternetService	DSL	DSL	DSL	DSL	Fiber optic
OnlineSecurity	No	Yes	Yes	Yes	No
OnlineBackup	Yes	No	Yes	No	No
DeviceProtection	No	Yes	No	Yes	No
TechSupport	No	No	No	Yes	No
StreamingTV	No	No	No	No	No
StreamingMovies	No	No	No	No	No
Contract	Month-to-month	One year	Month-to-month	One year	Month-to-month
PaperlessBilling	Yes	No	Yes	No	Yes
PaymentMethod	Electronic check	Mailed check	Mailed check	Bank transfer (automatic)	Electronic check
MonthlyCharges	29.85	56.95	53.85	42.3	70.7
TotalCharges	29.85	1889.5	108.15	1840.75	151.65
Churn	No	No	Yes	No	Yes

DATASET

- The dataset consists of 7043 rows and 21 features
- The feature “Churn” will be our label – “Yes” or “No”
- Some features description:
  - SeniorCitizen – whether customer is a senior citizen (0 or 1)
  - Partner – whether customer has partner (“Yes” or “No”)
  - Dependents – whether customer has dependents (“Yes” or “No”)
  - Tenure – number of months customer stay with the company
  - PhoneService – whether customer has phone service (“Yes” or “No”)

DATASET

➢ Some features description:

- MultipleLines – whether customer has multiple lines (“Yes”, “No” or “No phone service”)
- InternetService – customer internet service provider (“DSL”, “Fiber optic” or “No”)
- OnlineSecurity – whether customer has online sercurity (“Yes”, “No” or “No internet service”)
- OnlineBackup – whether customer has online backup (“Yes”, “No” or “No internet service”)
- DeviceProtection – whether customer has device protection (“Yes”, “No” or “No internet service”)

DATASET

➢ Some features description:

- TechSupport – whether customer has tech support (“Yes”, “No” or “No internet service”)
- StreamingTV – whether customer has streaming TV (“Yes”, “No” or “No internet service”)
- StreamingMovies – whether customer has streaming movies (“Yes”, “No” or “No internet service”)
- Contract – the customer’s contract term (“Month-to-month”, “One year”, “Two year”)
- PaperlessBilling – whether the customer has paperless billing (“Yes” or “No”)

DATASET

- Some features description:
  - PaymentMethod – customer’s payment method (“Electronic check”, “Mailed check”, “Bank transfer (automatic)”, “Credit card (automatic)”)
  - MonthlyCharges – customer’s monthly charges
  - TotalCharges – customer’s total charges over the period he/she stay
- Features that might be important are those that:
  - Describe types of services customers subcripe like “PhoneService”, “Multiplelines”, “ internet”.
  - Describe customer’s account information like ‘tenure”, “Contract”, “PaymentMethod”, “PaperlessBilling”, “MonthlyCharges” and “TotalCharges”

CHALLENGE

- Although we had pinpoint some features that might be importance, we still need to do an analysis (correlation, distribution) on all the features with the label to confirm
- To decide on the encoding method for categorical features
- If scaling is needed on the dataset
- Since "Churn" ("Yes" or "No") will be the label, our model will be a **Supervised Classification Model**. We have to try a few models from Logistic Regression, K-Nearest Neighbors, Decision Trees, Random Forest, Support Vector Machine, Naive Bayes and select thee best performed model for our problem

QUESTION TO ASK OURSELF

- Is there any useful features that we can create by using the existing features?
- Is there any missing values in the dataset? What is our approach if there is any?
- Is there a lot of outliers? How are we going to deal with it (remove or keep)?
- Do we want to drop those features that are not that useful?
- Is it an imbalance dataset? If it is, what metrics should be considered as the performance metric for our model?

THANK YOU