Telco-customer-churn-prediction

This is a classification machine learning problem for predicting customers churn from the company based on customers who left within the last month labeled by 'yes' or 'no'

The dataset used in this project includes information about:

customerID - Customer ID

gender - Whether the customer is a male or a female

SeniorCitizen - Whether the customer is a senior citizen or not (1, 0)

Partner - Whether the customer has a partner or not (Yes, No)

Dependents - Whether the customer has dependents or not (Yes, No)

tenure - Number of months the customer has stayed with the company

PhoneService - Whether the customer has a phone service or not (Yes, No)

MultipleLines - Whether the customer has multiple lines or not (Yes, No, No phone service)

InternetService - Customer’s internet service provider (DSL, Fiber optic, No)

OnlineSecurity - Whether the customer has online security or not (Yes, No, No internet service)

OnlineBackup - Whether the customer has online backup or not (Yes, No, No internet service)

DeviceProtection - Whether the customer has device protection or not (Yes, No, No internet service)

TechSupport - Whether the customer has tech support or not (Yes, No, No internet service)

StreamingTV - Whether the customer has streaming TV or not (Yes, No, No internet service)

StreamingMovies - Whether the customer has streaming movies or not (Yes, No, No internet service)

Contract - The contract term of the customer (Month-to-month, One year, Two year)

PaperlessBilling - Whether the customer has paperless billing or not (Yes, No)

PaymentMethod - The customer's payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))

MonthlyCharges - The amount charged to the customer monthly

TotalCharges - The total amount charged to the customer

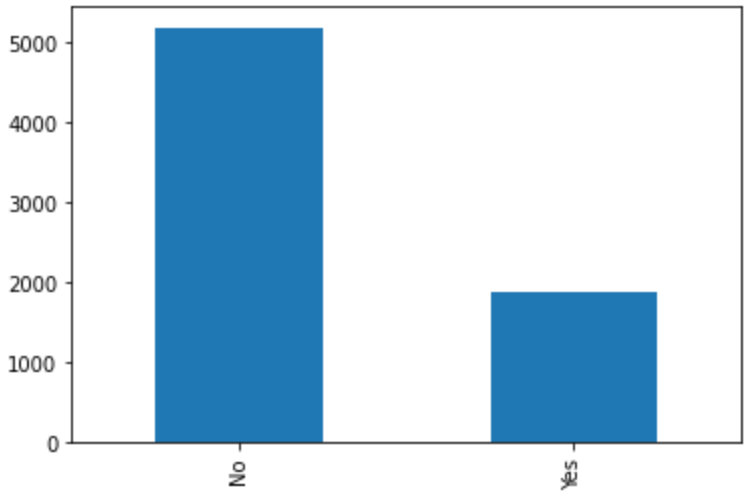
Churn - Whether the customer churned or not (Yes or No)

Data cleaning

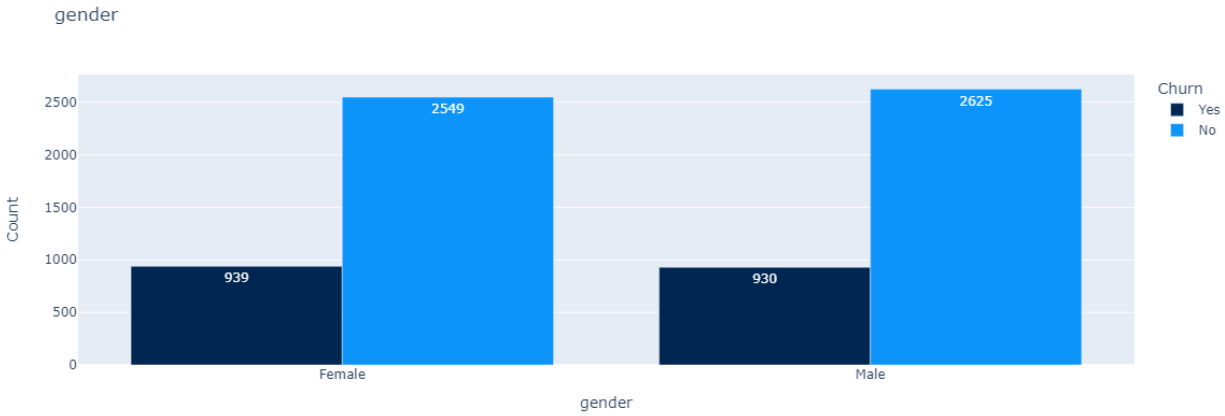
- At first, the dataset looks pretty clean with the right data types except for **TotalCharges** which has the data type `object` but it should be `float` this is because there are some blank spaces in that feature.
- So we check for similar things in other features too but other features look pretty clean.
- Finally, we replace the blank features with the median of that column as the distribution is skewed.

Exploratory data analysis

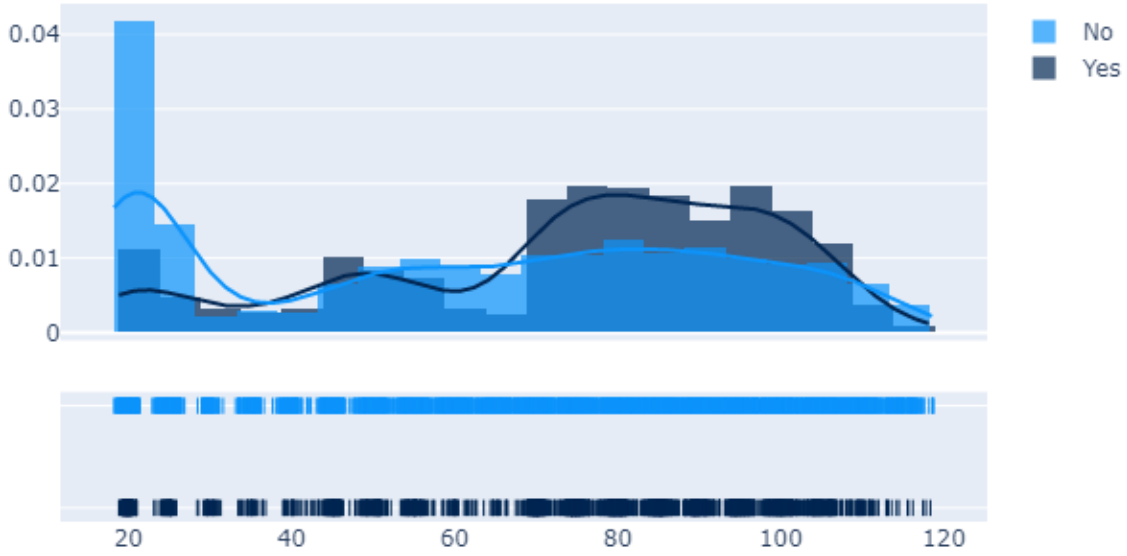
- 1st we check for unique features in each categorical variable.
- The dataset is highly skewed. About 73% for `No churn` and 27% for the `churn` category.



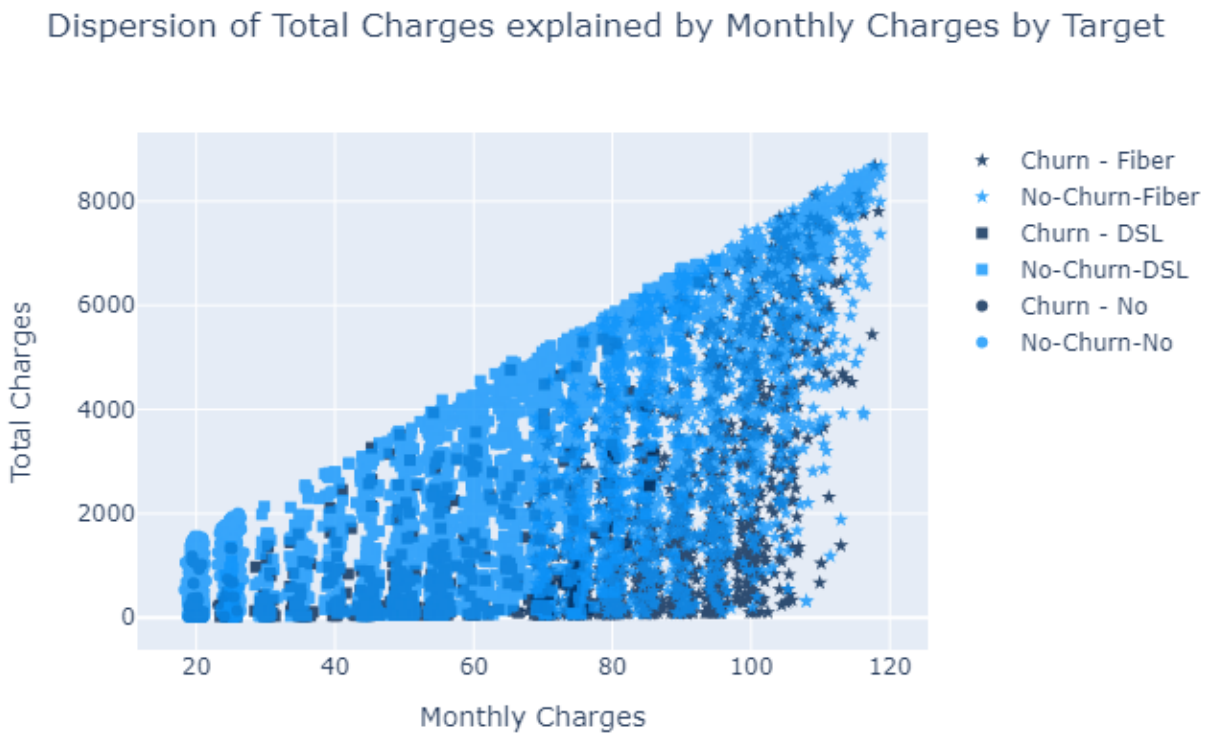
- Now we create a function to plot each categorical feature and see that most of the features are not uniform except for `gender`



- Then we plot the MonthlyCharges with churn yes & No. We observe that higher MonthlyCharhave has more chances of Churning compared to lower monthly charges.



- We see the opposite trend for TotalCharges**.**
- Then we plot a to scatter plot between total TotalCharges and MonthlyCharges and we can see a "linear function" where the two features are very correlated and it that makes a lot of sense.
  I thought that could be interesting if we divide the Total charges by the Monthly Charges and we will get the months till the Churn... It would be the very close value of tenure



All the plots are made using plotly here I have mentioned few of them so might not be visible in GitHub, for that you have to open and maybe run in jupyter notebook

Feature encoding

- I have encoded all the categorical features using label encoder

Feature engineering

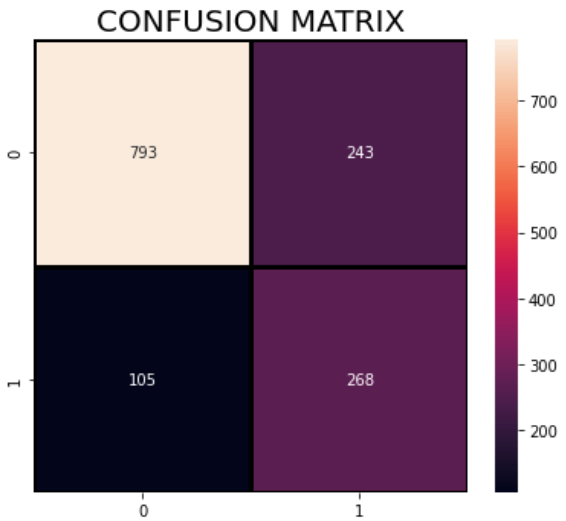
- Checked for correlation and as expected TotalCharges is highly correlated with MonthlyChargesso I dropped it.

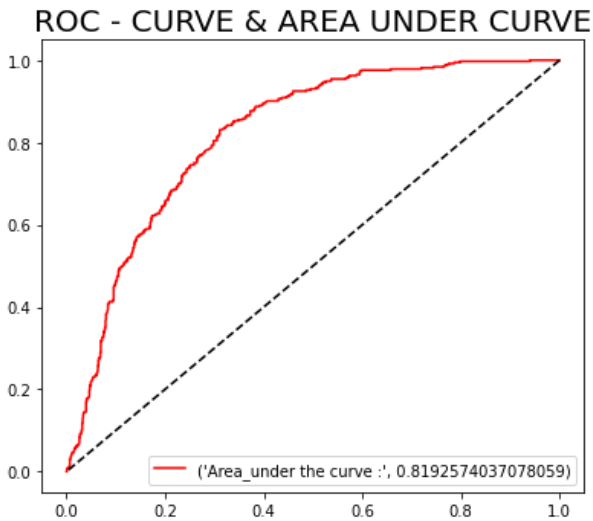
- Split the data into training and testing in an 8:2 ratio.
- As the dataset was highly imbalanced so I used SMOTE to balance the features.
- Now for deploying the best 6 features we used different feature selection methods.
- First I tried the variance threshold which checks for the features having a variance less than a given threshold, using this we remove the PhoneService.
- After that we use SequentialFeatureSelector to select 6 best features
- Then done the scaling and stored the data for further modeling.

Models training

Four different models weretopplied on the data and all results are reported with confusion matrix and classification report showing the precision, recall f1-score metrics and ROC curve.

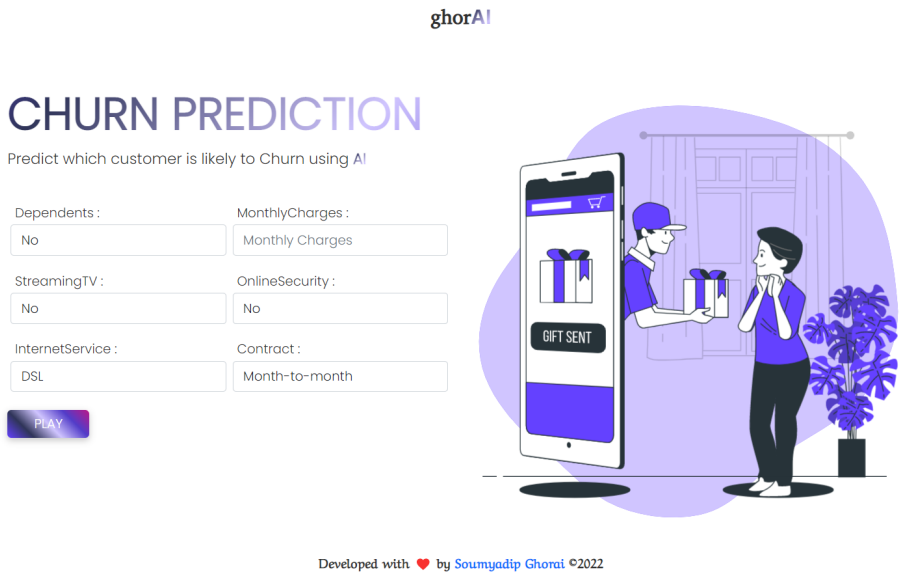
Here I have tried different algorithms like logistic regression, random forest, and XGBoost, SVC, done hyperparameter tuning for each of them and the above-mentioned metrices.
Finally I tried the stacked classifier model combining random forest, XGBoost and SVC which gave the best result with 75% accuracy so I proceed with that for deploying.





Deploy

For deploying I have used flask in the backend and and in the frontend I have used HTML, CSS and bootstrap.
After preparing the whole website I have deployed it using Heroku. This was we can reach to millions of users.



Future Work

- Future work might include storing the more features of users that might be a better indicator of their churn and work on those to reduce churning that can be done by giving discounts and offers
- So for that we can create an automatic pipeline to predict and giving offers. If we predict that the user might leave the service then we can start giving them offers and discounts thus they don't leave.

GitHub repo : https://github.com/soumyadipghorai/customer-churn

D**eployed link :** https://customer-churn-ml.herokuapp.com/