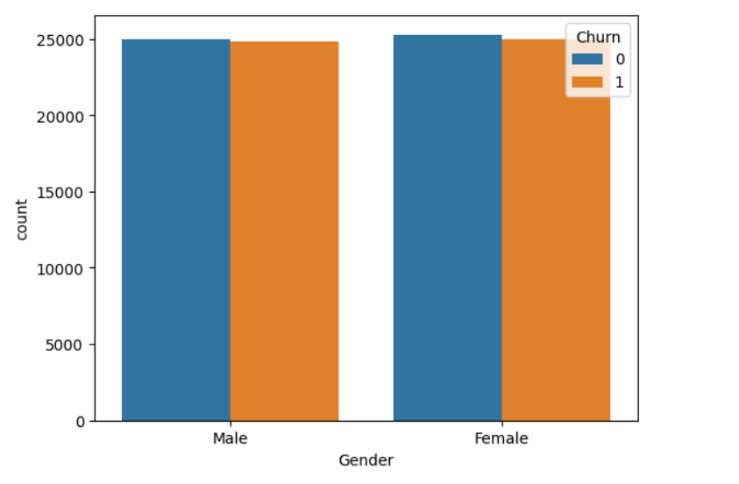
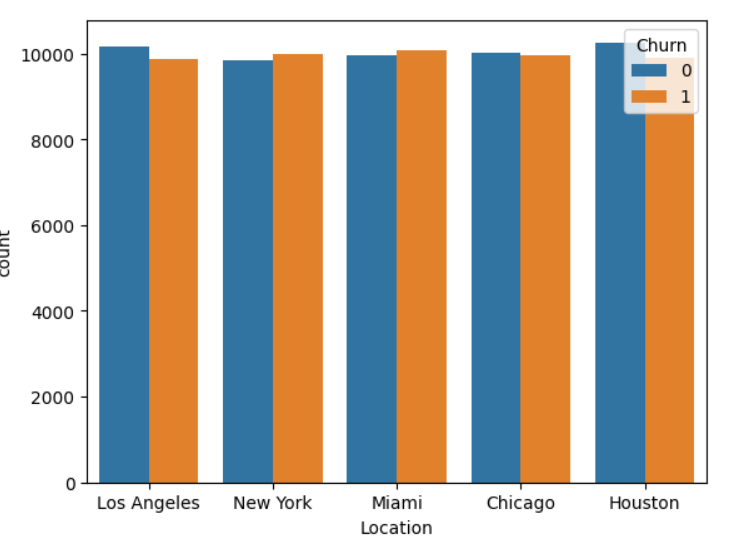
1. **Importing dataset and performing EDA and Feature engineering for data preprocessing and model selection**

On performing EDA on the provided dataset it was found that the input features were very minutely correlated to the customer churn and therefore didn’t play a significant role in customer churn prediction. This can be seen so by the various plots

i)**Gender distribution wrt to churn**

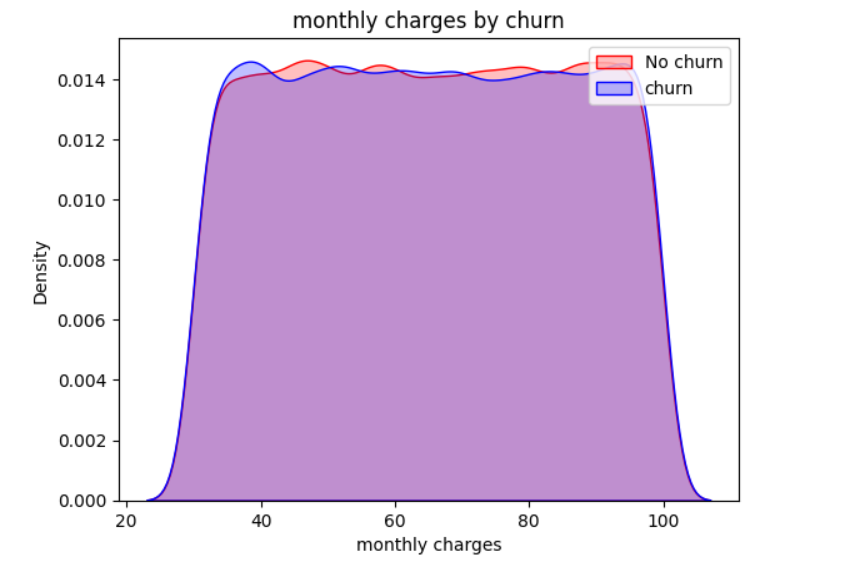
Both the male and female customers are nearly equal in case of churning therefore it isn’t possible to draw a significant conclusion about if gender individually plays a role in deciding wether a customer will churn on not .

**ii) Location and Churning**



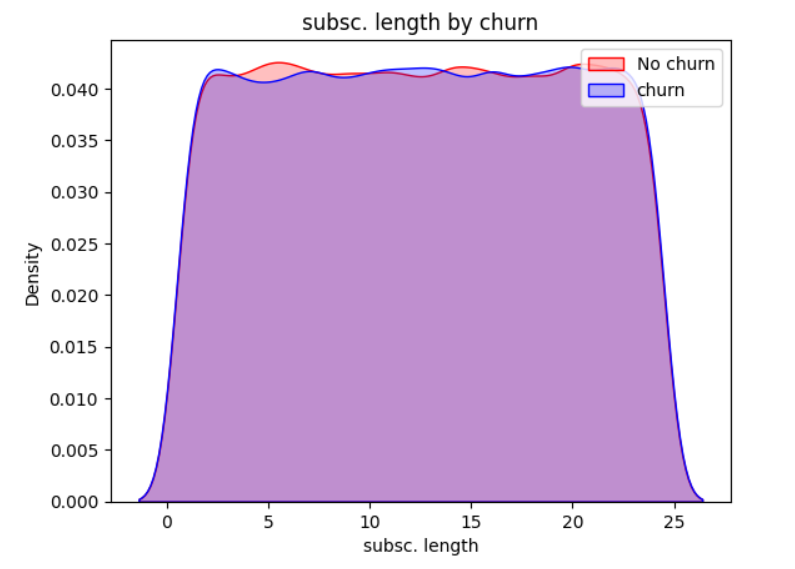
Just like gender location doesn’t carry a significant bias towards any specific region and all the locations are nearly equal in case of churning population , therefore location individually is not a strong indicator of churning population as well.

**iii) Monthly bill and churning**



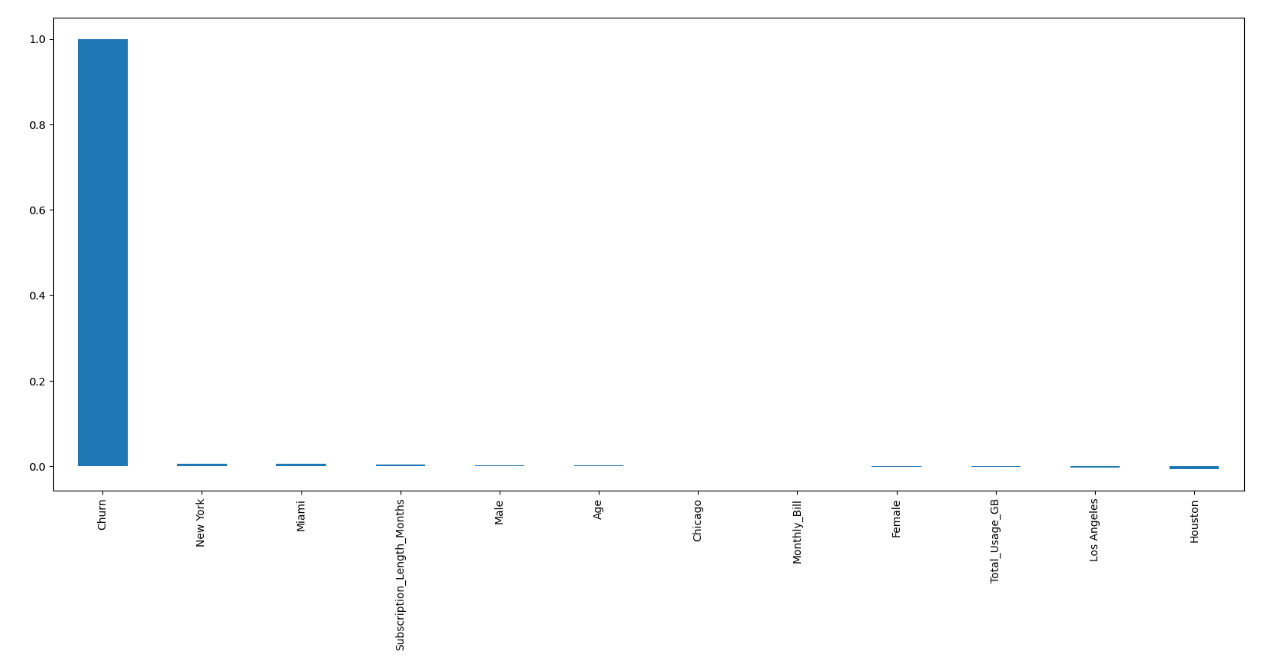
Here as well the monthly charges are not playing a significant role in deciding wether a customer will churn or not as we can see from the plot based on the data provided . The plot is a platue for a large range of monthly charges ,which signifies that the density of churners and non churners is almost similar for any given charge.

**iv)months of subscription**



Similar to the monthly charges , the duration of subscription is also not a strong determining factor for the same .

**v)correlation matrix**



From the confusion matrix we can see that the correlation between any given input feature and the churn is not very significant and therefore the input features / data are not very detrimental of the customer churn .



The above plot show the values of correlation between churn and other features as well as between the input features as well .

We can see the insignificant relation between churn and other input features which suggests that any input data would be almost un-useful for determining the customer churn and therefore the models will be predicting among 0 and 1 randomly which would give us around 50 percent accuracy which is justified by the math .

The probability of any of the two independent events to occur being 50 percent each .

\*\*\***Feature eng.** –

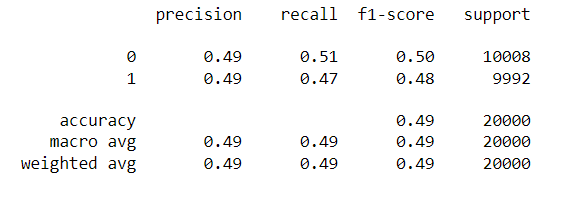
The categorical features ( age, location ) were converted into numerical one hot vectors using the panda’s get dummies function and the one hot encoder using sklearn and the whole dataset was scaled to reduce the variance in the data and reduce the deviation caused by outliers.

1. **Model selection**

Based on the conclusions from eda , I used 4 classification models and 1 custom made neural net implemented using pytorch and scikit learn.

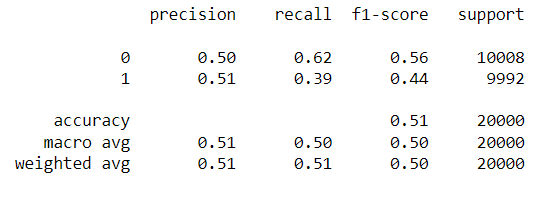
1. **RandomForestClassifier**

On using random forest classifier the following results were achieved



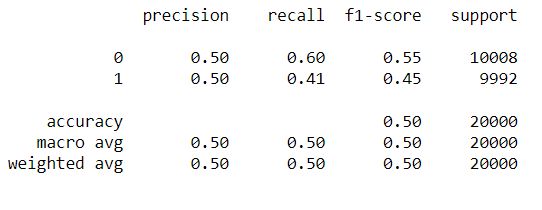
1. **LogisticRegression**

Logistic regression gave an accuracy and f1 score similar to random forest classifier

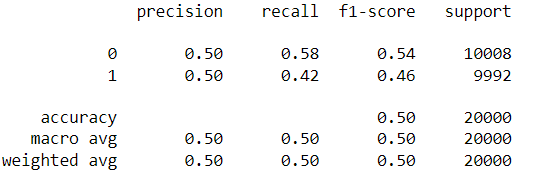


1. **Support vector machine.SVC**

The SVC classifier took a comparatively longer training and loading time and had an accuracy close to the previous two



1. **Gaussian naïve bayes**



1. **Custom neural network**

The nn architecture comprised of:

***3 pairs of alternate Linear and ReLU perceptrons*** for inducing non linearity to the network , in order to capture any latent relation between churn and given features

***The in\_features for the three pairs were taken as 11 and so were the out features***

For the last layer ***the in\_feature being 11 and outfeature being 2 to create 2 logits*** which would then be passed through the ***activation fuction ,softmax*** in this case , and then argmax of the generated 2 probabilities would give the y\_pred which would then be used to calculate the accuracy

Which came around to be 50 as well

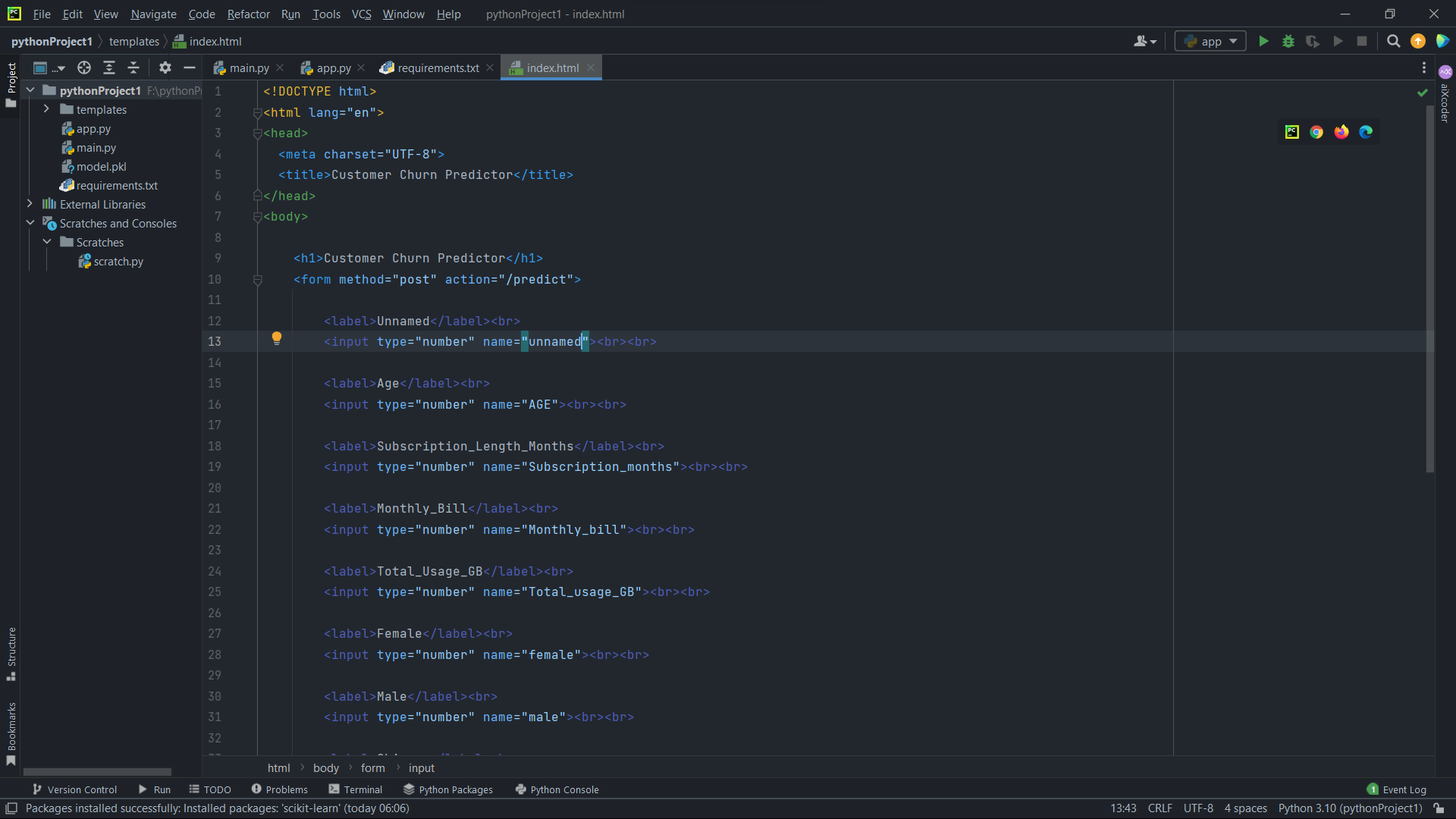
All this proves that the data points didn’t influence the customer churn as concluded from EDA .

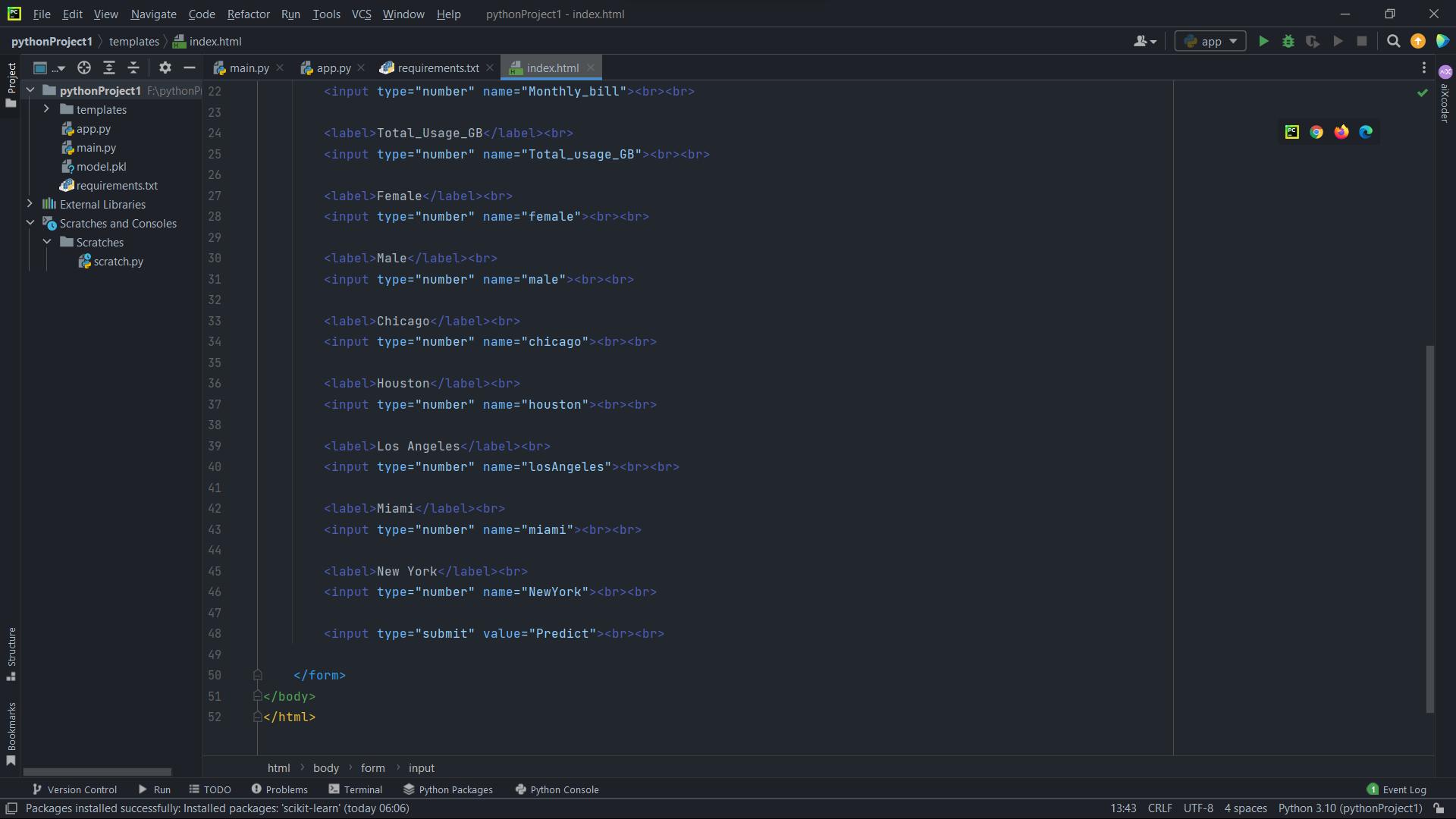
1. **Local site generation using Flask and HTML for deploying model and taking user input**

The site is a simple working model which takes in the same features as the dataset and predicts wether the customer will churn in future or not

The Flask+html local site file contains the

1. **Templates folder** – It includes the index file which is the HTML part of the local host site

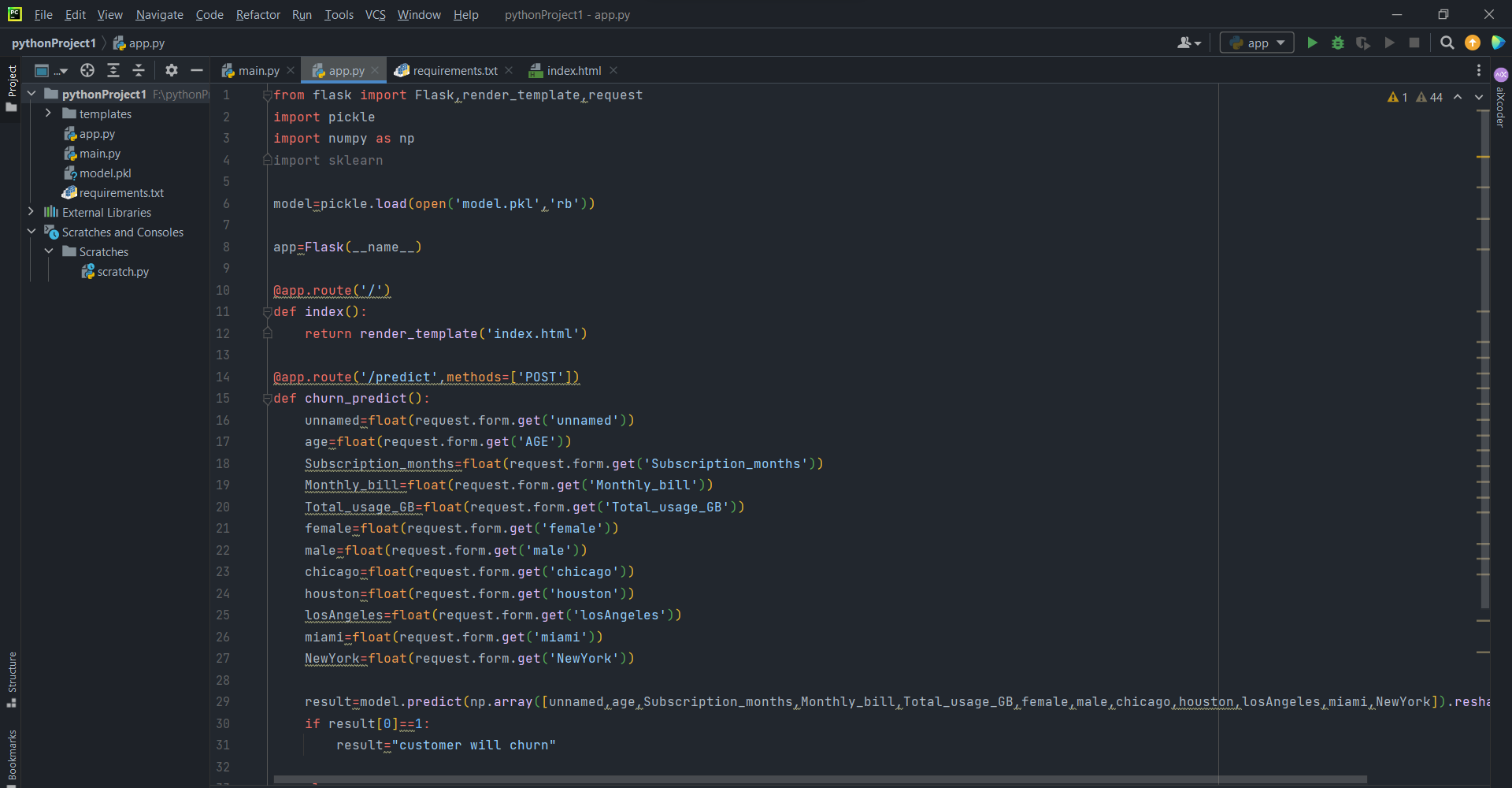


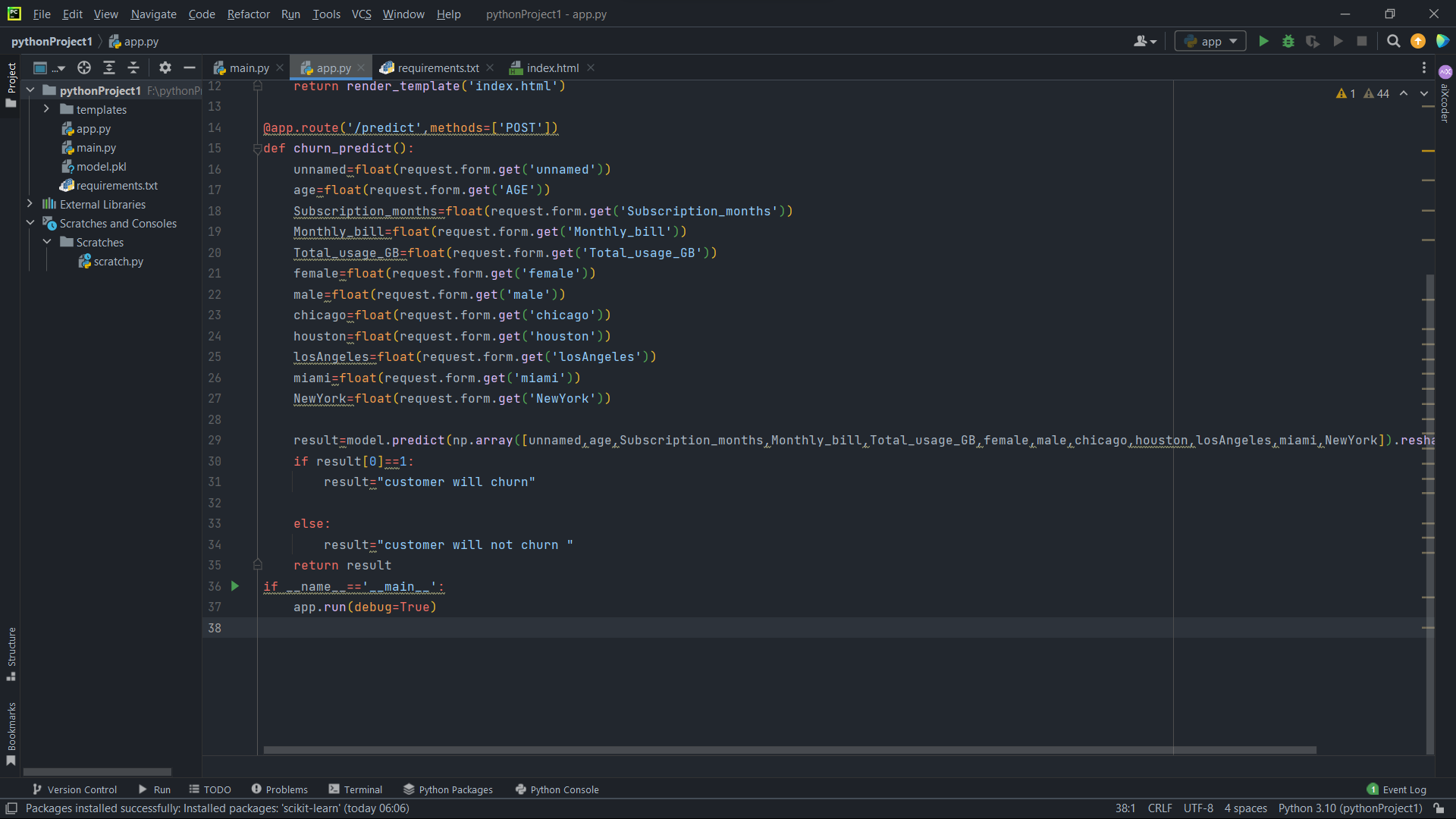


The above script is the code for the host html str.

1. **App.py** –

The app file contains the flask integrated code that call for the html body and has necessary imports and result returns which is shown as the output in the host site.





iii)**Model.pkl** –

The model.pkl file is the pickle format export of the final\_model jupyter notebook The pkl file contains the dataset created after performing eda and feature engineering to get numerical features and the model used is Logistic Regression as the f1 score achieved was the comparative highest amongst all the above mentioned models.