**Identifying which factors predict future user adoption**

**Data Cleaning**

The columns were labeled as adopted user or not using the user\_engagement data frame. The ratio of the classes was 87:13 with the adopted users as the minority class. The time difference between creation time and last login was calculated by calculating the time delta. The first step was to convert the date columns to the same date time unit objects. The creation time was converted to unix timestamp in seconds and the difference was calculated. An additional column was added to indicate whether a member was invited or not. The creation source column was one-hot encoded.

**Model**

The XGBoost regression model was the best model of all the models that were trained on the data. The performance metrics were significantly better in comparison. The hyperparameters are shown in table 3. The model performed similarly for both for the train and test data. Thus this model can be reliably used for predicting cancer mortality rate using the regional socioeconomic variables indicated in this model.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **XGB** | **Precision** | **Recall** | **F1** | **Accuracy** | **ROC** |
| **Class 0** | 1 | 0.95 | 0.97 | 0.95 | 0.99 |
| **Class 1** | 0.74 | 0.97 | 0.84 |  |  |

Table 2. Performance metrics of the XGB regression model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **XGB**  **Parameter** | **learning**  **\_rate** | **max\_**  **depth** | **scale\_pos\_weight** | **n\_estimators** |
| **Best** | 0.01 | 3 | 6 | 500 |

Table 3. Hyperparameters for the best XGB regression model.

The built-in feature importance function was used to score the features based on their weighted contribution to the model. The scores showed that the time between creation of account and the last login was the most important factor in determining user adoption. Signing up using google or by being invited were minor factors.