CSE 523: Machine Learning

Group 17 - Hardly Humans
Weekly Project Report - 5

Quora Insincere Questions Classification

Name	Enrolment Number
Malav Doshi	AU1940017
Parth Shah	AU1940065
Sanya Zaveri	AU1920064
Mihir Pathak	AU1920138

1) Tasks Performed in the week.

- Feature engineering
- Modeling
- Inferences

2) Outcomes of the tasks performed.

- Adding additional features to the raw data using domain knowledge.
- These extra features help improve the quality of results.

```
data.head()
```

qid	question_text	target	num_words	num_unique_words	len_text	len_char_question_text	len_word_qu
00002165364db923c7e6	How did Quebec nationalists see their province...	0.0	13	13	72	24	13
000032939017120e6e44	Do you have an adopted dog, how would you enco...	0.0	16	15	81	20	16
0000412ca6e4628ce2cf	Why does velocity affect time? Does velocity a...	0.0	10	8	67	20	10
000042bf85aa498cd78e	How did Otto von Guericke used the Magdeburg h...	0.0	9	9	57	24	9

- Selecting logistic regression model for training data.

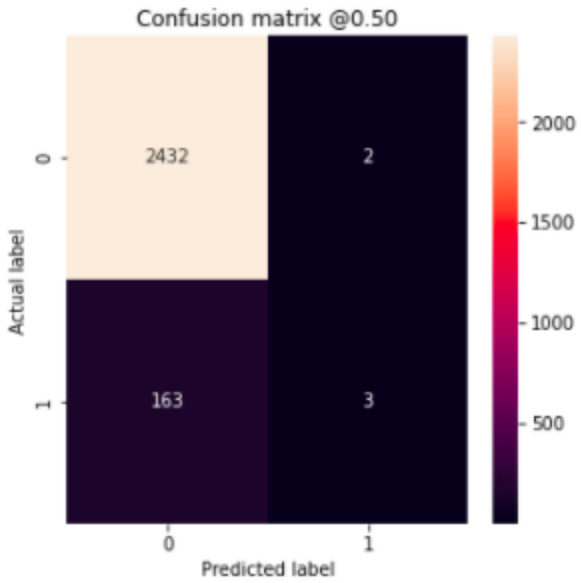
```
y = train.iloc[:,2].values
columns = [3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19]
X = data.iloc[0:1306122,columns].values
X_train, X_test, y_train, y_test = \
    train_test_split(X, y, test_size=0.20, random_state=42)
import psutil
psutil.virtual_memory()

logres = linear_model.LogisticRegression(C=0.1, solver='sag', max_iter=1000)
logres.fit(X_train, y_train)

lr_preds = logres.predict(X_test)
from sklearn.metrics import f1_score
F1_score = f1_score(y_test, lr_preds, average='weighted')
print("Logistic regression F1 score: %0.3f" % F1_score)
```

```
Logistic regression F1 score: 0.912
```

- Finding accuracy, precision, recall and f1 score.
- Implementing the confusion matrix.



3) Tasks to be performed in the upcoming week.

- Dividing dataset into train, validation and test and performing model selection.
- Looking for other ML algorithms like Naïve Bayes and SVM.
