2019

Craigslist housing scams in San Francisco

MGMT 590-049 | GROUP 2

LIYAO GAO SR| MANKARAN SINGH BAHRI | RUKMINI SUNIL NAIR | RUNQI GE |
SAGAR KURADA | XEMA VINOD PATHAK

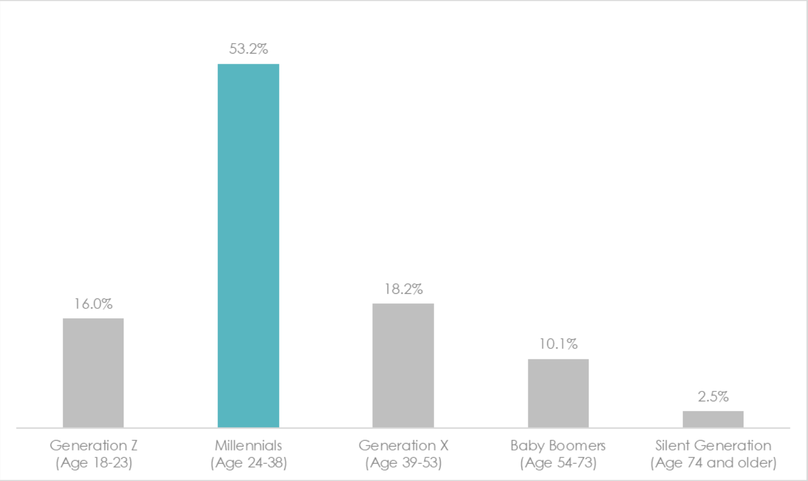
Introduction

The percentage of services that are conducted online is increasing rapidly with increasing internet penetration. One such service industry is online apartment rentals. As people gain access to the internet and get faster connections, more people are expected to use online resources when finding apartments.

Industry outlook

The online apartment rental services industry is at a relatively nascent stage and has grown rapidly over the five years to 2019 at an annual growth rate of 15.7%. As people have become increasingly comfortable using the internet for their house-hunting process, the industry demand has skyrocketed. As per the 2019 IBIS report, the industry revenue stands at $446.2 million with a profit of $50 million. In the coming five years, it is projected that this industry will grow at an annualized rate of 10.7%, garnering a revenue of $740.7 million.

Online rental companies operate listing websites that enable house-hunters to search for apartments based on a variety of criteria, such as neighborhood, amenities offered, number of bedrooms and bathrooms and price. These websites showcase the listing with detailed description and images of the property that helps the customers make an informed decision. Millennials (age 24-38) constitute 53.2% of the customer-base of such online rental websites and 99% millennials use the internet to look for apartments. Zillow, rent.com, and Craigslist are major players in the market for the online rental industry.

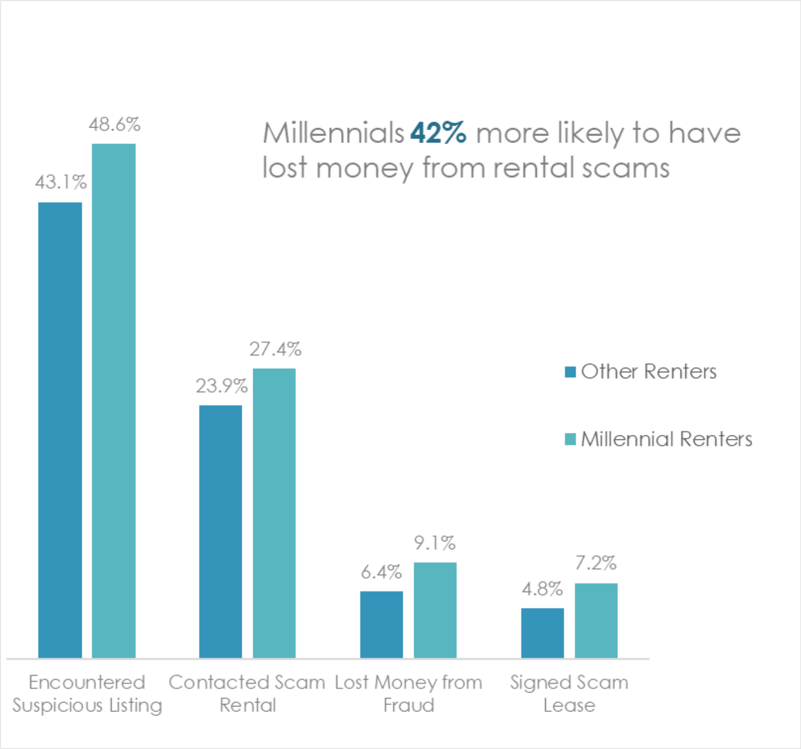


What is Craigslist?

Started as an email distribution list to friends in San Francisco, Craigslist became one of the biggest American classified advertisements websites with various sections such as jobs, housing, services, gigs and many more. Craig Newmark in 1993 started Craigslist as an email service to let people know about events happening in and around the San Francisco Bay Area. Through word of mouth, the number of subscribers exploded and this service transitioned into an online classified service with millions of listings for various categories. The site gets more than 20 billion pageviews and 49.4 million unique visitors every month, ranking at 9th place overall among websites in the United States. With more than 80 million new classified advertisements each month, Craigslist is the leading classifieds service in any medium. Craigslist also dominates the U.S. rental housing market, with millions of new listings every month. The 23 largest U.S. cities listed on the Craigslist home page collectively receive more than 300,000 postings per day just in the "for sale" and "housing" sections. However, website popularity is prone to many scam incidents.

Online rental scams in the USA

With the growing market for the online rental industry, the number of fraudulent incidents has also seen a significant rise. As per the Apartment Fraud Survey, 43.1% of the house-hunters encountered a fraudulent listing. US renters lost about $5.2 million because of fraudulent apartment listing. As most of the millennials resort to online services for their apartment rental needs, they are highly prone to encountering the fraudulent listings. Of the renters who lost money, 1 in 3 lost more than $1000. Out of all the cities having a presence in the online rental industry, San Francisco (47.8%) tops the chart of the percentage of renters who encountered fraudulent listing followed by Los Angeles (46.7%).



As per the study conducted by the United States Census Bureau, an average American will move 11.7 times in the entire lifetime, leading to the continuous growth of the online rental industry. With the exponentially rising listings for online rentals on its website, Craigslist is struggling with keeping the website free from scammers. It is estimated that about 6% of the housing listings are fraudulent, resulting in the loss of millions of dollars.

Business Objective

It is a known fact that the success of Craigslist is limited by persistent scams. The possibility of being scammed deter potential customers from using the platform, resulting in lower average revenue per customer. The impact of Craigslist on the economy is not non-negligible as such advertisements could reduce the housing rental vacancy rates. The expected value of loss due to a scam listing can be thought of as a probable tariff towards future transactions. Keeping the positive effects of Craigslist towards the economy in the center, it is imperative for the product managers at Craigslist to come up with a clean and credible platform.

The main objective of this project is to help product managers at Craigslist achieve the goal of user satisfaction. It is evident from secondary research that millennials are the primary target audience of the website and are most vulnerable to online frauds. Empathizing with millennial house-hunters, it can be said that they are the ones looking for legitimate advertisements of housing rentals on online platforms so that they could easily find a nice and affordable place near their workplace.

To achieve this goal, it is necessary to flag a suspicious ad listing on the website and take necessary action against it. Our approach constitutes a mix of primary and secondary research for feature selection and model formation.

Related work

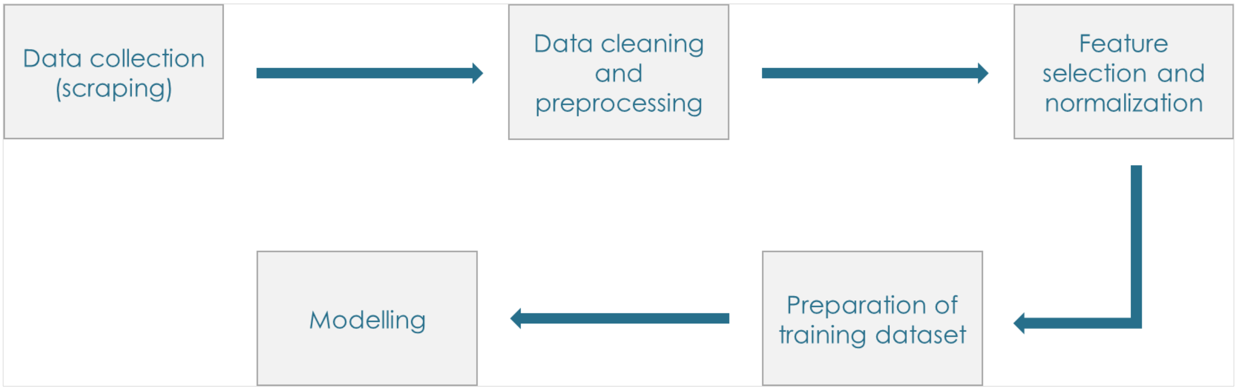
Much of the prior research has identified content-based features as key identifiers for any ad listing to be spam. These features include the title, description, and images of the listing.

The spammers use templates to automatically generate ads and post them on various websites. The ads can be identified by studying various patterns in the text such as character and word count, sentence construction, special characters and selection of words. One such research presents a comprehensive taxonomy of webspam practices. Another research focused on the relevance of the images for spam detection. An irrelevant image typically constitutes of alphanumeric characters or links, suggesting that the ad listing has a higher probability of being a scam.

The content-based features can be combined with the numerical features such as price, number of bedrooms, number of bathrooms, square feet area, number of images et al. to improve the quality of detection.

Due to a lack of publicly available training dataset, the training dataset for this project was built based on this research.

Framework


Data collection (scraping)
Data cleaning and preprocessing
Feature selection and normalization
Preparation of training dataset
Modelling

We started by collecting data for training and test datasets from Craigslist.com for a list of exhaustive features derived from the secondary research. Post that, we flagged the training dataset to classify a listing as spam/ not spam based on the patterns concluded in various papers published studying the spam listings. The numerical features were then normalized to make a consolidated training dataset. This dataset was then used to build a classification model for the rental listings.

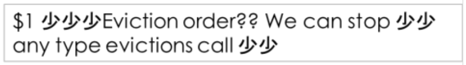
Data collection and analysis

For the purpose of modeling, a training dataset was formed as a combination of content-based and numerical features. Text and image analytics was performed on unstructured data, and the numerical variables were normalized.

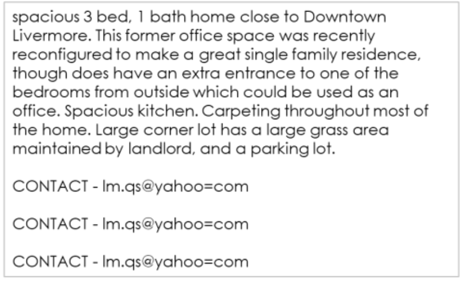
Text analytics

A sample of 3000 titles and descriptions for various locations were analyzed to identify spam patterns. As per the taxonomy of webspam, the following parameters were considered to flag a suspicious listing:

- Asking for personal information
- Fishy-looking e-mail address or domain
- Beautiful unit for pennies on the dollar
- No security deposit, a month’s free rent


$1 少少少Eviction order?? We can stop 少少 any type evictions call 少少

Example of scam title


spacious 3 bed, 1 bath home close to Downtown Livermore. This former office space was recently reconfigured to make a great single family residence, though does have an extra entrance to one of the bedrooms from outside which could be used as an office. Spacious kitchen. Carpeting throughout most of the home. Large corner lot has a large grass area maintained by landlord, and a parking lot.

CONTACT - lm.qs@yahoo=com

CONTACT - lm.qs@yahoo=com

CONTACT - lm.qs@yahoo=com

Example of scam description

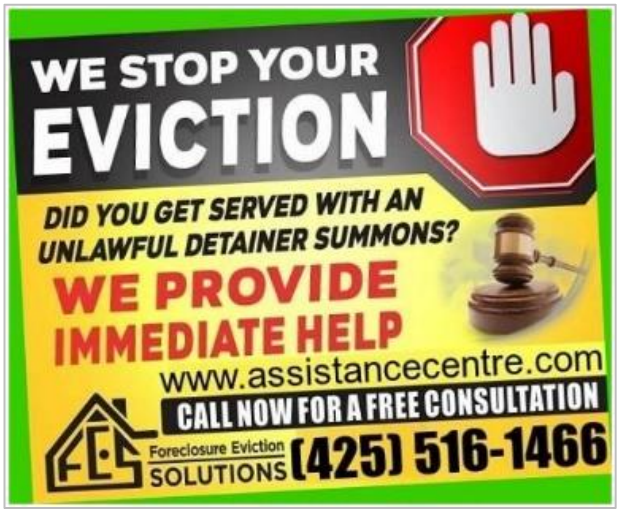
- Irrelevant or unrelated information
- Poorly constructed sentences, excessive capitalization

Process:

1. Tokenization, lemmatization, removal of punctuations and stop-words
2. TFIDF vectorization
3. Apply the Naïve Bayes and Neural Networks model for text analytics

For the final text analytics model, the Neural networks model was selected as it resulted in higher validation accuracy.

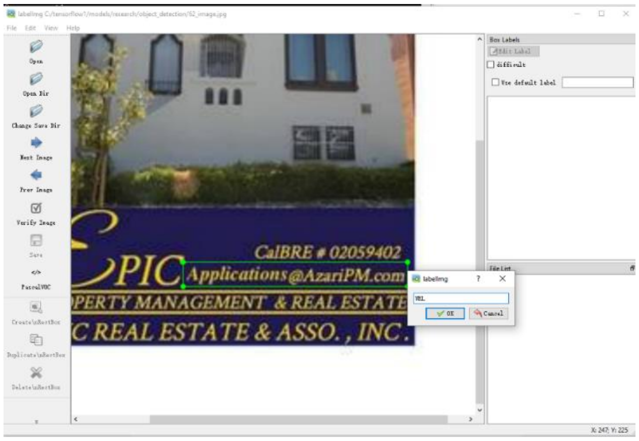
Image analytics



Listings having images carrying text, phone numbers or email addresses are usually suspicious. This model uses TensorFlow to identify such images and flags them if any alphanumeric characters are spotted.

TensorFlow provides a collection of detection models pre-trained on the COCO dataset, the Kitti dataset, the Open Images dataset, the AVA v2.1 dataset, and the iNaturalist Species Detection Dataset. These models can be useful for out-of-the-box inference if some categories already have been in those datasets. They are also useful for initializing our models when training on novel datasets. In our case, we choose faster_rcnn_inception_v2_coco.

Process:



1. **Setup:** Set up all required software about TensorFlow and set Anaconda virtual environment.
2. **Gather and label pictures:** We download over 300 images from craigslist which have text in their images. Through LabelImg package, draw a box around each object in each image and name the label.
3. **Generate training data:** Generate the TFRecords that serve as input data to the TensorFlow training model. Our model uses the xml_to_csv.py and generate_tfrecord.py scripts from Dat Tran's Raccoon Detector dataset, with some slight modifications to work with our directory structure.
4. **Training:** Create a label map and configure training.
5. **Run the training:** This process took about 10496 steps to get the loss lower than 0.05.
6. **Threshold:** If the model recognition probability is higher than 0.6, we flag the image as spam.

Other features

Apart from content-based features, we also considered various numerical and binary features for the final model based on secondary research.

Numerical:

1. Price of the property
2. Number of bedrooms
3. Number of bathrooms
4. Word count of the title
5. Word count of the description
6. Number of images in the listing

Binary:

1. Location: 1 if the neighborhood (location of the property) exists, else 0
2. Phone number: 1 if there is a phone number in the description text, else 0

3. Image analytics: 1 if the image is suspicious (as per the above-mentioned analysis), else 0

Preparation of consolidated training dataset

After building models for text and image analytics, they were applied to the final training dataset to get the numeric interpretation of the unstructured data. All the other numerical variables were normalized.

Normalization process:

The goal of normalization is to make every datapoint have the same scale so each feature is equally important. Here the scalar variables have different ranges and the difference in the ranges is also significant. Hence, when the algorithm compares data points, the feature with the larger scale will completely dominate the other. For this model, min-max normalization was performed.

Data modeling

Once the training dataset was developed, it was trained on various models.

1. **Train-test split:** The training dataset was split in a ratio of 75:25 for training and validation purposes.
2. **Models:** The classification models such as Logistic regression, Support Vector Machine, Random Forest and Deep Learning were applied to the training dataset.

Model validation

The statistical scam online rental advertisement detection model developed needs to be validated to estimate the accuracy with which it will predict the test data. A cross-validation method was used on the training and validation datasets to evaluate the prediction of the classification of an independent sample. We evaluated the models based on the accuracy and confusion matrix to decide upon the best model.

Model	**Accuracy**
Logistic Regression	95.04%
Naïve Bayes	69.6%
Support Vector Machine	98.08%

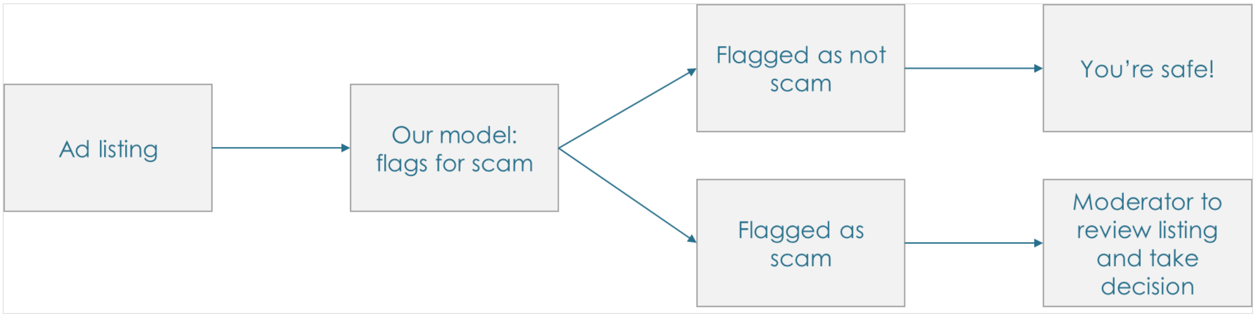
Random Forest	86.72%
Deep Learning	99.36%

**Please refer appendix for the model implementation*

The accuracy of the Deep Learning model with 2 layers and 3 neurons was the best and hence that was used for the test data. As we increase the training dataset for the unstructured part and include more conditions to determine to scam posts, the accuracy would further drop.

Conclusion

This project provides a solution to make the online rental section of Craigslist spam-free to ensure higher customer satisfaction. Once the model is applied to the platform, the way ahead for the company would be to review the suspicious ad listings and take necessary actions.


Ad listing
Our model: flags for scam
Flagged as not scam
You're safe!
Flagged as scam
Moderator to review listing and take decision

Benefits

- **Risk mitigation:** As highlighted in the introduction, 33% of the people who encountered fraudulent listings lose more than $1000. The process of online house hunting is perceived as highly risky because of difficulty in evaluating the authenticity of the listing. Hence, an automated system for fraud detection can make the house-hunters less vulnerable to financial loss.
- **Increased security:** One of the prominent patterns in the scam description is asking for personal information, such as SSN, bank details, drivers' license, upfront money for the deposit and giving a free stay. Online forums such as Reddit host discussions on such cases. It was observed on these platforms that there have been plenty of incidents of identity theft and bank fraud due to disclosure of personal details. This model aims to detect such patterns in the listings and flag those ads for a further review, ensuring customer security.

- **Increased customer satisfaction:** The customer satisfaction score is dependent on the ease of use of the platform and the relevance of search results. This model will flag and filter out irrelevant listings, resulting in high customer satisfaction. The customer can find relevant ads quickly and easily, leading to high customer lifetime value on the website and recurring usage.
- **Increased credibility:** Craigslist has increasingly become prone to the “bait-and-switch” phenomenon, negatively impacting the credibility of the platform. This model could further be tweaked for other listings for scam detection and improve the quality of listings rest of the categories. An increased customer satisfaction fetches positive word-of-mouth, resulting in increased credibility of the platform.

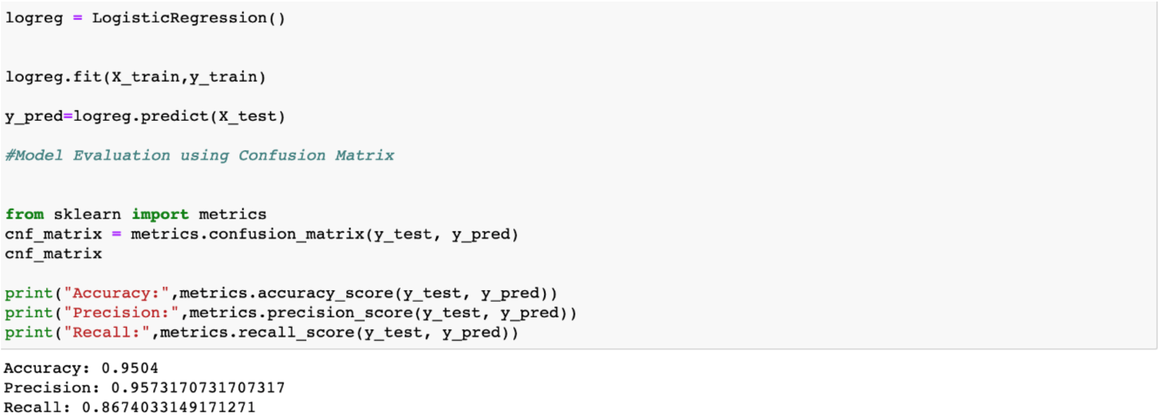
Future enhancements

- **Categorization by neighborhood:** This model treats the entire San Francisco area as same. However different neighborhoods have different rents for a given number of bedrooms, bathrooms and square feet area. Another level of sophistication can be added to this model by studying various neighborhoods separately to make a robust prediction for rent variable.
- **Special characters in text:** The peer-reviewed research papers on webspam suggest a correlation between high usage of special characters and spam. An advanced text analytics model could be developed by incorporating the pattern of special character usage to detect a scam listing.
- **Language compatibility:** This model was trained for the text in the English language only. It could be further enhanced by incorporating text analytics in various other languages. This will enable to expand across geographies, ensuring a cleaner international online classified platform.
- **Enhanced image analytics:** This model only considered the presence of alphanumeric characters and links on the images to classify suspicious listings. However, a more enhanced model could be trained on housing and non-housing images to ensure that the images are relevant to the listing. A housing image could incorporate images of neighborhood, society, images of property rooms (bedroom, living room, kitchen, bathroom).

Appendix

Models results

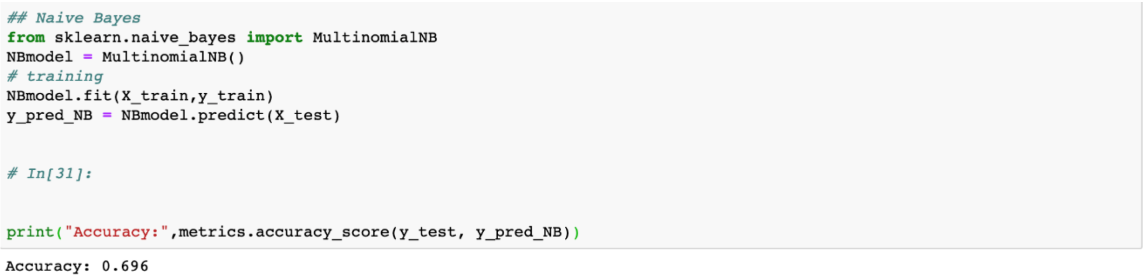
Logistic Regression



```
logreg = LogisticRegression()


logreg.fit(X_train,y_train)

y_pred=logreg.predict(X_test)

#Model Evaluation using Confusion Matrix


from sklearn import metrics
cnf_matrix = metrics.confusion_matrix(y_test, y_pred)
cnf_matrix

print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
print("Precision:",metrics.precision_score(y_test, y_pred))
print("Recall:",metrics.recall_score(y_test, y_pred))
```

```
Accuracy: 0.9504
Precision: 0.9573170731707317
Recall: 0.8674033149171271
```

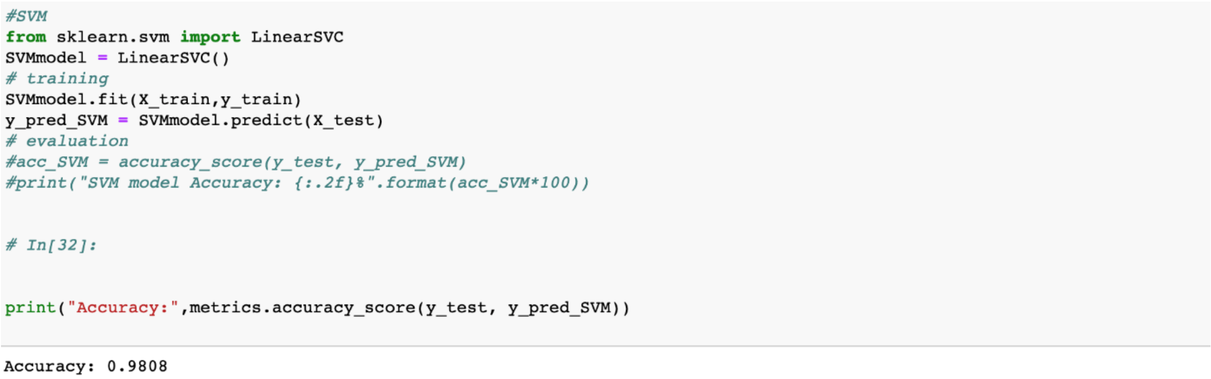
Naïve Bayes



```
## Naive Bayes
from sklearn.naive_bayes import MultinomialNB
NBmodel = MultinomialNB()
# training
NBmodel.fit(X_train,y_train)
y_pred_NB = NBmodel.predict(X_test)


# In[31]:


print("Accuracy:",metrics.accuracy_score(y_test, y_pred_NB))
```

```
Accuracy: 0.696
```

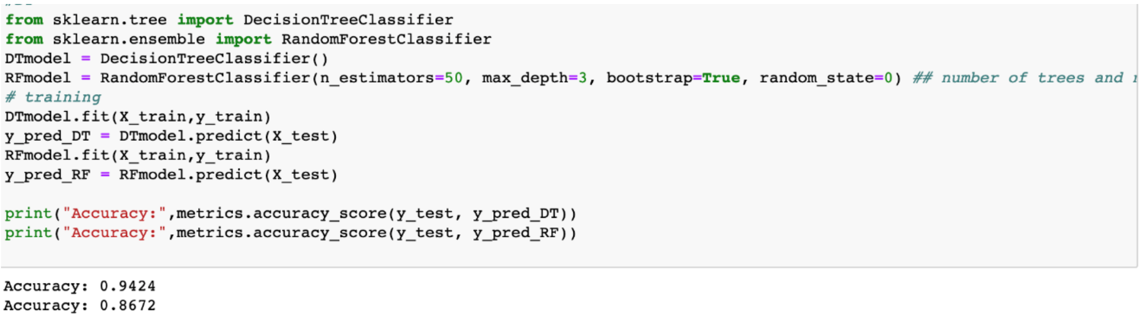
Support Vector Machine



```
#SVM
from sklearn.svm import LinearSVC
SVMmodel = LinearSVC()
# training
SVMmodel.fit(X_train,y_train)
y_pred_SVM = SVMmodel.predict(X_test)
# evaluation
#acc_SVM = accuracy_score(y_test, y_pred_SVM)
#print("SVM model Accuracy: {:.2f}%".format(acc_SVM*100))


# In[32]:


print("Accuracy:",metrics.accuracy_score(y_test, y_pred_SVM))
```

```
Accuracy: 0.9808
```

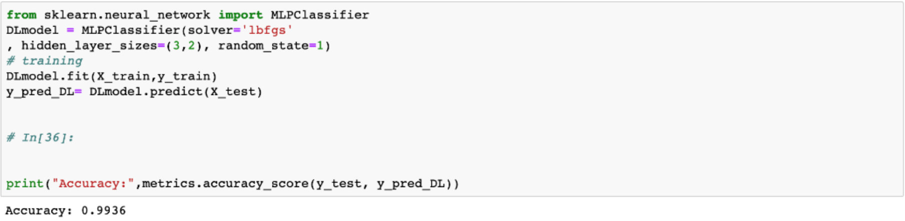
Random Forest Classifier



```
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
DTmodel = DecisionTreeClassifier()
RFmodel = RandomForestClassifier(n_estimators=50, max_depth=3, bootstrap=True, random_state=0) ## number of trees and
# training
DTmodel.fit(X_train,y_train)
y_pred_DT = DTmodel.predict(X_test)
RFmodel.fit(X_train,y_train)
y_pred_RF = RFmodel.predict(X_test)

print("Accuracy:",metrics.accuracy_score(y_test, y_pred_DT))
print("Accuracy:",metrics.accuracy_score(y_test, y_pred_RF))
```

```
Accuracy: 0.9424
Accuracy: 0.8672
```

Deep Learning



```
from sklearn.neural_network import MLPClassifier
DLmodel = MLPClassifier(solver='lbfgs'
, hidden_layer_sizes=(3,2), random_state=1)
# training
DLmodel.fit(X_train,y_train)
y_pred_DL= DLmodel.predict(X_test)


# In[36]:


print("Accuracy:",metrics.accuracy_score(y_test, y_pred_DL))
```

```
Accuracy: 0.9936
```

References:

Cook, D. (2019). For lease: Increased internet usage has fueled substantial industry growth. *IBIS World*.

Bennet, S., & Popov, I. (2018, July 20). Million Dollar Scam: Rental Fraud Costs 5.2 Million U.S. Renters. Retrieved from https://www.apartmentlist.com/rentonomics/how-common-is-rental-fraud-scams/

Z. Gyongyi and H. Garcia-Molina. Web spam taxonomy. *In First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb 2005)*, April 2005.

B. Mehta, S. Nangia, M. Gupta, and W. Nejdl. Detecting image spam using visual features and near duplicate detection. *In Proceeding of the 17th international conference on World Wide Web, WWW '08*, pages 497–506, New York, NY, USA, 2008. ACM

Tran, H., Hornbeck, T., Ha-Thuc, V., Cremer, J., & Srinivasan, P. (2011). Spam detection in online classified advertisements. *Proceedings of the 2011 Joint WICOW/AIRWeb Workshop on Web Quality - WebQuality 11*. DOI: 10.1145/1964114.1964122