Complaints Text Classification

UDACITY MACHINE LEARNING ENGINEER NANODEGREE

Sérgio da Costa
November 14th, 2020

Abstract

With the evolution of technologies there has been a growing demand for Machine Learning (ML) systems. Companies are increasingly looking for these systems to enhance their products and solve business problems. From my professional experience, and from all the organizations I've transited, there are always classification problems, namely text. Problems related to product categorization or mapping, support ticket classification, analysis and tagging of data collected from UX Research area surveys, etc. Due to the impossibility to use data from a real problem from my organization, I had to choose a real data subset of complaints received about financial products and services that will allow me to explore some supervised ML techniques for multi-class text classification. The aim is to have a system that allows free text complaints to be classified automatically to a product /or with a single click, and that allow to analyse which products are more frequent at the same time. In this project, the aim was to use ML methods to model such a classifier. It investigated the relationship between products and complaints, and compared various algorithms such as Logistic Regression, Support Vector Machines, Random Forest, Naive Bayes Classifier, K-Nearest Neighbors and Decision Trees, using cross validation to obtain the accuracy of each model. I identified the two best ML models for the problem: Linear SVC and Logistic Regression models, using the bag of words from TF-IDF, and I have tried to explore these two models, using validation and prediction techniques. In the end, the web app was developed to interact with the model.

Keywords: *Machine Learning, Multi-Class, Text Classification, Logistics Regression, SVM, Linear SVC, Complaints*

Definition

1. Project Overview

With the exponential growth of volume and diversity of data available in organizations, with cheaper, more powerful, faster and more effective processing, with most accessible storage, we have seen a growth in the demand for Machine Learning (ML) solutions as it's possible to produce more fast and automatic models capable of analyzing complex and huge volumes of data while delivering faster and more accurate results.
One of the core functions in ML is to get computers to automatically find a good predictor based on past experiences. This kind of work is done through classification. Classification is the process of using a model to predict unknown values (output variables), using a number of known values (input variables) [1].

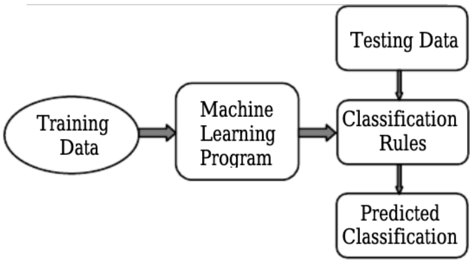


Figure 1 - Classification Architecture [1].

In classification we may have unsupervised approaches when the instances (input) are not labelled or supervised approaches when they are given with known labels (i.e., the corresponding correct results).

Classification is the task of choosing the correct class label for a given input. In basic classification tasks, each input is considered in isolation from all other inputs, and the set of labels is defined in advance. Some examples of classification tasks are [1]:

- Identify if an email is spam or not.
- Deciding what the topic of a news article is, from a fixed list of topic areas such as "sports," "technology," and "politics."

The basic classification task has a number of interesting variants. For example, in multi-class classification, each instance may be assigned multiple labels; in open-class classification, the set of labels is not defined in advance; and in sequence classification, a list of inputs is jointly classified.

When we talk about organizing text, we are basically saying that we intend to organize it into certain subject topics, or specific categories. There are several examples where we can find this organization, the products of a company are organized by category or

family of products, customers can be segmented into several categories based on a set of relevant features, or the categorization of tickets that report problems. On the other hand, we have binary classification problems such as sentiment analysis, bank fraud, forum moderation, moderation of classified or e-commerce sites, email spam identification, among many other examples [2].
With the use of ML models, in particular Text classification models, we are able to solve several problems.

In this context, the main objective of this project is to identify the best model to classify complaints about financial products and services [3] [4] from customers into already predefined classes. For this I will evaluate several ML models [1] [4] [5] and identify the best performance so that I can automate a solution that automatically classifies all the complaints that arise on a daily basis.

I will use a dataset called *Consumer Complaint Database (Consumer Finance Complaints - Bureau of Consumer Financial Protection)* which can be downloaded from Kaggle [6] or Data.gov [7]. These data are real world complaints received about financial products and services "*Each complaint has been labeled with a specific product; therefore, this is a supervised text classification problem. With the aim to classify future complaints based on its content, we used different ML algorithms can make more accurate predictions (i.e., classify the complaint in one of the product categories). The dataset contains different information of complaints that customers have made about a multiple products and services in the financial sector, such us Credit Reports, Student Loans, Money Transfer, etc. The date of each complaint ranges from November 2011 to May 2019*" [6].

It's important to mention that my main motivation for this project is to work with several supervised ML models to solve text classification or multi-class problems. This is because over the last few years I have had several classification problems such as mapping generic products with a specific catalogue, processing data collected in UX research areas through surveys, in several languages, for a possible classification in specific tags predefined in the organization, or even make the sentiment analysis of these responses, or classification in 3 hierarchical levels of categories of the tickets we collect daily in Zendesk, etc. I would be happy to work on a set of real data from the organization where I work, but for confidentiality reasons that is not possible. So, as an alternative, I chose to select the dataset I refer to below, and the goal is to work on these ML classification concepts and new algorithms and find solutions that can be applied on a professional level.

In the end, I will be able to have a model capable of being applied in this particular context, with considerable accuracy, or even of replicating it with the appropriate changes, to the professional problem.

2. Problem Statement

Based on the selected dataset I was faced with a supervised text classification problem, as there is a classification base already in place. The main goal is to understand within the various ML models for classification, which are the most appropriate to solve this problem.

And why do I want to solve this problem?

Besides my personal motivation presented above, being in a production environment, every time a new product or financial service complaint comes in, the idea is to automatically classify this complaint to one of the existing products (categories), without the need to do the manual classification, often time-consuming and requiring human resources to do it, a boring task that after a few days leads to demotivation.

The problem to be solved is classification task using supervised learning (Figure 2): The best supervised model of classification (with a high accuracy) to automatically classify only one product for each complaint.

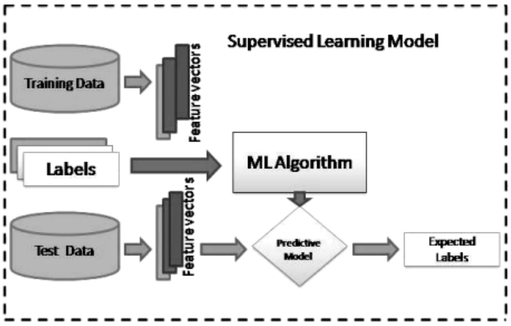


Figure 2 - Supervised ML to a real-world problem [1].

The comparison between supervised ML methods can be made by making statistical comparisons of the accuracy of trained classifiers in specific datasets [1].

Considering some of the models that the authors presents [1] [8] [9] I will evaluate the following algorithms:

- Support Vector Machines (SVM) (Linear SVC)
- Logistic Regression.
- Random Forest.
- Naïve Bayes Classifier – Gaussian / Multinomial.
- K-Nearest Neighbors.
- Decision Trees.

I applied different models on the same test dataset and the same training dataset.

The performance evaluation of the models will be based on the metrics presented in the evaluation metrics section.

The approach proposed here is to train one or two supervised models to classify the complaints that enter daily. Once I have trained and identified the best models, I can automatically classify all the complaints in the training set and then I can compare features of a certain classification with the training classifications.
In the end I will have a model capable of predicting the best product for the complaints that will enter the system in the future.

3. Evaluation Metrics

Selecting an algorithm to get the best results is an important step in an ML project, like the one I have at hand.
The evaluation of an algorithm is often judged by its accuracy. The evaluation of algorithms is most often based on the accuracy of the prediction and can be measured using the following metric [1]:

***Accuracy**: Number of correct classifications / Total number of test cases*

Being a supervised, classification problem, Accuracy will be one of the metrics I will use to evaluate the performance of the models [19]. It will make sense to use this metric if there is a uniform distribution of samples across existing classes, in order not to induce an error. To minimize the margin of error, and achieve a better prediction, if this happens the stratification method is used. This is the process of reorganizing the data to ensure that each fold is a good representative of the whole.

However, the idea is also to evaluate other metrics that can help to understand the performance of the models:

1. Classification report for each label/class to evaluate the metrics:

- **Recall**: True positives / (True positives + False negatives)
- **Precision**: True positives / (True positives + False positives)
- **F1-Score:** will be high when both metrics are high and similar

I also considered the use of the **Confusion Matrix** and **Cross-Validation** important to evaluate the models.

2. The Confusion Matrix will show the number of cases in which model was right or wrong in each product.

3. Cross-validation is one of the most useful techniques to evaluate different combinations of feature selection, dimensionality reduction, and learning algorithms [1], [20]. I chose the StratifiedKFold method, which is a variation of KFold [28]. First, StratifiedKFold shuffles the data and then splits the data into n_split parts (figure 3). This option is due to the fact that there are unbalanced product classes, i.e., part of the classifications are in may number in some specific classes.

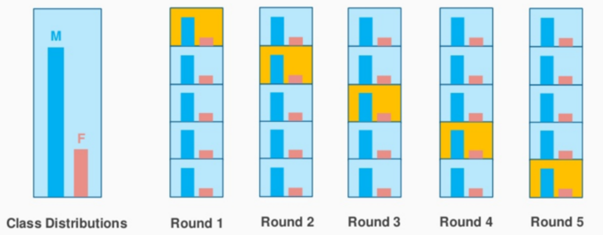


Figure 3 - Example of 5 folds Stratified Cross Validation [28].

In the project I will explore these metrics in order to validate the best model for our problem.

4. Project Design Methodology

In literature there are several frameworks that support us in the process of understanding, designing and implementing ML models in an organized, cyclical and structured way [21] [22] [23].
Text classification is an example of supervised ML task since a labelled dataset containing text documents and their labels is used for training a classifier. An end-to-end text classification pipeline is composed of three main components [22] [24]:

1. *Dataset Preparation*
2. *Feature Engineering*
3. *Model Training*

Considering this, I chose to define the following methodology:

- **Prepare environment**: the first step is to prepare the Python + Jupyter Notebooks environment to run the various models.
- **Dataset preparation:** the second step includes the process of loading the dataset and performing basic data exploration and pre-processing. The dataset is then splitted into train and validation sets.

- **Feature engineering:** it’s in this step that the raw dataset is transformed into flat features which can be used in a ML model (creating new features from the existing data).

- **Model training:** the final step is the model building step in which a ML model is trained on a labelled dataset.

- **Improve performance:** based on the results, look at different ways to improve the performance of text classifiers. Finding and applying the best hyperparameters [26].

- **Choosing the best model** for the text classification problem.

- **Deployment model**: based on the selected model, ensure an application or tool for viewing and interacting complaints vs. products.

At any time, I can return to an earlier step to refine.

Analyze

1. Data Collection

After understanding the problem at hand, I proceeded with the first step of collecting and loading the data into our platform.
I collected and stored complaints data so that they can be worked on in a next stage of exploration

2. Data Exploration

The first step in working with the data set is to load the data and identify what information is included in the data set. I prepared a notebook to do all the data exploration and patterns observation on the features given to it and the data distribution, as well as the definition of my final data set, to be worked on in the next step of pre-processing and data feature engineering.

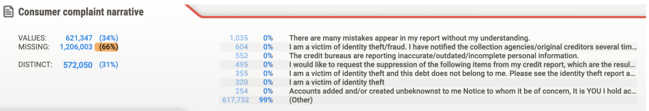
With resources from different Exploratory Data Analysis (EDA) and data pre-processing methodologies and libraries, I explored and visualized this data from the existing data set and analyse the possible metrics that will be used to understand the solution to the problem:

- A general statistical summary of the data set, using EDA visualization tools.
- Distribution of complaints, product, sub product, issue and sub issue.
- Relationship of complaint text length and assigned product.
- Relationship of the complaint narrative to the product allocated.
- Relationship of the words of the complaint to the product.
- Try to identify the relationship between features.
- Reduce our data subset.

In a first analysis, the data set has a structure of 18 columns and about 1.8 million records. This is a considerable data set.

I started by generating an EDA report in order to understand the data and its correlations. Using Sweetviz I generated an HTML report (i.e., Figure 4 and Figure 5), through which it was possible to perform a high-level analysis.

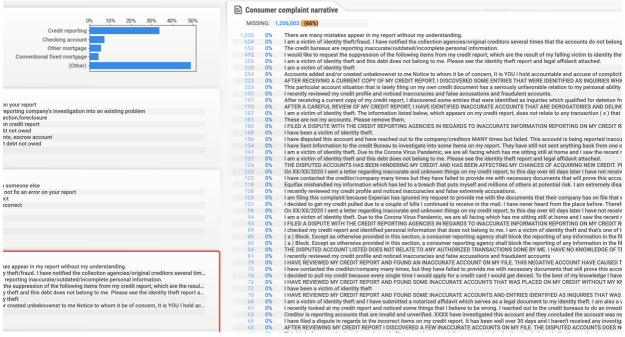
Considering the objective of the project, the analysis of the data set will focus on two attributes: the complaint narratives and the associated products.



```
Consumer complaint narrative

VALUES:     621,347   (34%)        1,035    0%   There are many mistakes appear in my report without my understanding.
MISSING:  1,206,003   (66%)          604    0%   I am a victim of identity theft/fraud. I have notified the collection agencies/original creditors several tim...
                                     552    0%   The credit bureaus are reporting inaccurate/outdated/incomplete personal information.
DISTINCT:   572,050   (31%)          495    0%   I would like to request the suppression of the following items from my credit report, which are the resul...
                                     355    0%   I am a victim of identity theft and this debt does not belong to me. Please see the identity theft report a...
                                     320    0%   I am a victim of identity theft
                                     254    0%   Accounts added and/or created unbeknownst to me Notice to whom it be of concern, It is YOU I hold ac...
                                 617,732   99%   (Other)
```

Figure 4 – EDA for complaints narrative text.



```
on your report
reporting company's investigation into an existing problem
ection,foreclosure
on credit report
bt not owed
nts, escrow account
t debt not owed

someone else
not fix an error on your report
ect
ncorrect

kes appear in my report without my understanding.
y theft/fraud. I have notified the collection agencies/original creditors several tim...
reporting inaccurate/outdated/incomplete personal information.
the suppression of the following items from my credit report, which are the resul...
y theft and this debt does not belong to me. Please see the identity theft report a...
y theft
r created unbeknownst to me Notice to whom it be of concern, It is YOU I hold ac...
```

```
Consumer complaint narrative
MISSING:  1,206,003 (66%)

1,035   0%   There are many mistakes appear in my report without my understanding.
  604   0%   I am a victim of identity theft/fraud. I have notified the collection agencies/original creditors several times that the accounts do not belong
  552   0%   The credit bureaus are reporting inaccurate/outdated/incomplete personal information.
  495   0%   I would like to request the suppression of the following items from my credit report, which are the result of my falling victim to identity the
  355   0%   I am a victim of identity theft and this debt does not belong to me. Please see the identity theft report and legal affidavit attached.
  320   0%   I am a victim of identity theft
  254   0%   Accounts added and/or created unbeknownst to me Notice to whom it be of concern, It is YOU I hold accountable and accuse of complicit
  223   0%   AFTER RECEIVING A CURRENT COPY OF MY CREDIT REPORT, I DISCOVERED SOME ENTRIES THAT WERE IDENITIFIED AS INQUIRIES WHI
  203   0%   This particular account situation that is lately filing on my own credit document has a seriously unfavorable relation to my personal ability
  197   0%   I recently reviewed my credit profile and noticed inaccuracies and false accusations and fraudulent accounts.
  197   0%   After receiving a current copy of my credit report, I discovered some entries that were identified as inquiries which qualified for deletion fr
  193   0%   AFTER A CAREFUL REVIEW OF MY CREDIT REPORT, I HAVE IDENTIFIED INACCURATE ACCOUNTS THAT ARE DEROGATORIES AND DELINC
  187   0%   I am a victim of identity theft. The information listed below, which appears on my credit report, does not relate to any transaction ( s ) that
  181   0%   These are not my accounts. Please remove them.
  180   0%   I FILED A DISPUTE WITH THE CREDIT REPORTING AGENCIES IN REGARDS TO INACCURATE INFORMATION REPORTING ON MY CREDIT R
  168   0%   I have been a victim of identity theft.
  156   0%   I have disputed this account and have reached out to the company/creditors MANY times but failed. This account is being reported inaccu
  154   0%   I have Sent Information to the credit Bureau to investigate into some items on my report. They have still not sent anything back from one o
  147   0%   I am a victim of identity theft. Due to the Corona Virus Pandemic, we are all facing which has me sitting still at home and I saw the recent r
  137   0%   I am a victim of identity theft and this debt does not belong to me. Please see the identity theft report and legal affidavit attached.
  124   0%   THE DISPUTED ACCOUNTS HAS BEEN HINDERING MY CREDIT AND HAS BEEN AFFECTING MY CHANCES OF ACQUIRING NEW CREDIT. P
  123   0%   On XX/XX/2020 I sent a letter regarding inaccurate and unknown things on my credit report, to this day over 60 days later I have not receiv
  122   0%   I have contacted the creditor/company many times but they have failed to provide me with necessary documents that will prove this accou
  118   0%   Equifax mishandled my information which has led to a breach that puts myself and millions of others at potential risk. I am extremely disa
  108   0%   I recently reviewed my credit profile and noticed inaccuracies and false extremely accusations.
  103   0%   I am filing this complaint because Experian has ignored my request to provide me with the documents that their company has on file that v
  101   0%   I decided to get my credit pulled due to a couple of bills I continued to receive in the mail. I have never heard from the place before. Theref
   98   0%   On XX/XX/2020 I sent a letter regarding inaccurate and unknown things on my credit report, to this day over 60 days later I have not receiv
   94   0%   I am a victim of identity theft. Due to the Corona Virus Pandemic, we are all facing which has me sitting still at home and I saw the recent r
   93   0%   I FILED A DISPUTE WITH THE CREDIT REPORTING AGENCIES IN REGARDS TO INACCURATE INFORMATION REPORTING ON MY CREDIT R
   89   0%   I checked my credit report and identified personal information that does not belong to me. I am a victim of identity theft and that's one of t
   86   0%   ( a ) Block. Except as otherwise provided in this section, a consumer reporting agency shall block the reporting of any information in the fi
   85   0%   ( a ) Block. Except as otherwise provided in this section, a consumer reporting agency shall block the reporting of any information in the fi
   84   0%   THE DISPUTED ACCOUNT LISTED DOES NOT RELATE TO ANY AUTHORIZED TRANSACTIONS DONE BY ME. I HAVE NO KNOWLEDGE OF T
   81   0%   I recently reviewed my credit profile and noticed inaccuracies and false accusations and fraudulent accounts
   79   0%   I HAVE REVIEWED MY CREDIT REPORT AND FOUND AN INACCURATE ACCOUNT ON MY FILE. THIS NEGATIVE ACCOUNT HAVE CAUSED T
   75   0%   I have contacted the creditor/company many times, but they have failed to provide me with necessary documents that will prove this accou
   74   0%   I decided to pull my credit because every single time I would apply for a credit card I would get denied. To the best of my knowledge I have
   72   0%   I HAVE REVIEWED MY CREDIT REPORT AND FOUND SOME INACCURATE ACCOUNTS THAT WAS PLACED ON MY CREDIT WITHOUT MY K
   72   0%   I have been a victim of identity theft
   70   0%   I HAVE REVIEWED MY CREDIT REPORT AND FOUND SOME INACCURATE ACCOUNTS AND ENTRIES IDENTIFIED AS INQUIRIES THAT WAS
   68   0%   I am a victim of identity theft and I have submitted a notarized affidavit which serves as a legal document to my identity theft. I am also a v
   67   0%   I recently looked at my credit report and noticed some things that I believe to be wrong. I reached out to the credit bureaus to do an investi
   65   0%   Creditor is reporting accounts that are invalid and unverified. XXXX have investigated this account and they concluded the account was no
   65   0%   I have filed a dispute in regards to the incorrect items on my credit report. It has been well over 30 days and I haven't received any investig
   63   0%   AFTER REVIEWING MY CREDIT REPORT I DISCOVERED A FEW INACCURATE ACCOUNTS ON MY FILE. THE DISPUTED ACCOUNTS DOES N
```

Figure 5 – EDA details for complaints narrative text.

I filtered and excluded all complaints that did not have a narrative (null values), because beyond the space they occupied on the date set, they would muddle the model.

From 1.8 million complaints, there are about 600,000 cases with text (~ 34% of the original dataset is not null). This is still a good number to work with.

After this, taking into account the problem I have proposed, I selected only the relevant columns and renamed them, considering a new nomenclature.

3. Exploratory Visualization

Half of the complaint text is shorter than 707 characters. Though there are some very long pieces of text, 90 percent is shorter than 2304 and 95 percent is shorter than 3227 (Figure 6).

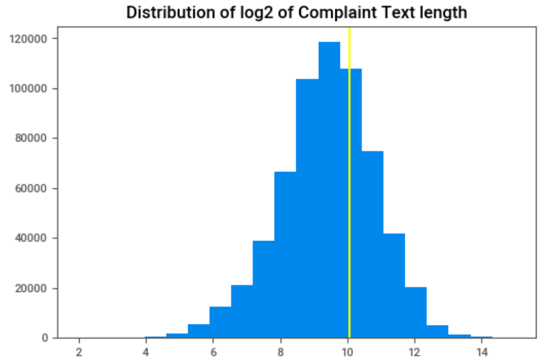


Figure 6 –Complaint narrative text distribution.

Since our classification label is the Product, it's important to understand its distribution and balancing.

When we have certain high levels of balancing, that is, when a large % of the complaints are rated for a reduced number of products, we may have some problems in the accuracy of the models (it can be high for this reason).

Below, you should notice two things (Figure 7):

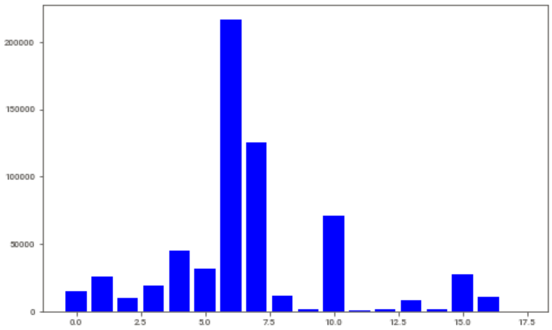


Figure 7 – Products distribution (before consolidation).

Through the analysis I find that the number of complaints per product is unbalanced. Consumer complaints are more focused on credit reporting, debt collection and mortgages, for instance. There are 18 different product classes (Figure - 8). However, it's observed that some classes are contained in others. For instance, 'Credit card' and 'Prepaid card' are contained in 'Credit card or prepaid card' products.

Here I tried to better balance these products, I consolidated some of these products (Figure 9), joining the products with multiple or repeated descriptions (there are several products of Loan, Credit Card, Prepaid, Payday, etc.).

Even so, I will have classes with more preponderance, which may even be reasonable as they may allow us a greater accuracy.

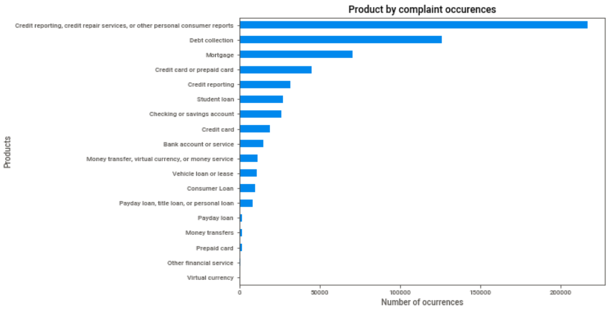


Figure 8 – Products distribution (18 products).

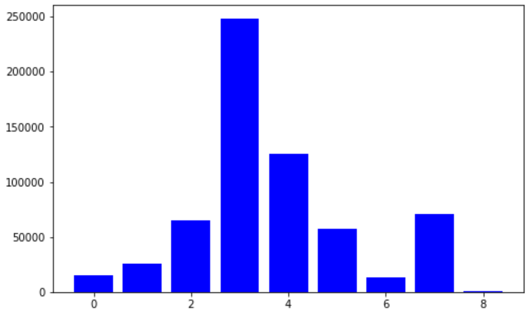
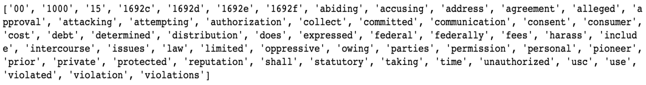


Figure 9 –Products distribution (before consolidation).

When analysing examples of narratives of the complaints I realise that it is complicated to read, because there are lots of characters and there are texts too long. This text can be worked on using pre-processing techniques. We can rely on stop words removal using Term Frequency–Inverse Document Frequency (TF-IDF).
Through the tests, it removes the punctuation and also all stop words (Figure 10). The spell checker can also confuse some words, thus changing the semantic meaning of the word.



```
['00', '1000', '15', '1692c', '1692d', '1692e', '1692f', 'abiding', 'accusing', 'address', 'agreement', 'alleged', 'a
pproval', 'attacking', 'attempting', 'authorization', 'collect', 'committed', 'communication', 'consent', 'consumer',
'cost', 'debt', 'determined', 'distribution', 'does', 'expressed', 'federal', 'federally', 'fees', 'harass', 'includ
e', 'intercourse', 'issues', 'law', 'limited', 'oppressive', 'owing', 'parties', 'permission', 'personal', 'pioneer',
'prior', 'private', 'protected', 'reputation', 'shall', 'statutory', 'taking', 'time', 'unauthorized', 'usc', 'use',
'violated', 'violation', 'violations']
```

Figure 10 –Example of TfidfVectorizer

I decided to clean up the text of the complaint narrative in the next phase (Text Processing) for better accuracy of the algorithms: converting all letters to lower. A few rules applied:

- Convert all letters to lower.
- Remove white spaces.
- Expand some abbreviations.
- Remove non ascii.
- Remove wrong conversion characters from text.
- Remove xxx since it will be treated as important words.

4. Algorithms and Techniques

Because it's a classification problem, the following algorithms have been used to predict the classification of a product regarding a complaint narrative text:

Support Vector Machines (Linear SVC): are a set of supervised learning methods which have been used for classification. The benefits of using SVM are: i) It is effective is high dimensional space, ii) Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient, iii) It is versatile because holds different kernel functions that can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels [1]. Effective in high dimensional spaces. Still effective in cases where the number of dimensions is greater than the number of samples. Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient [13].

Naive Bayes Classifier (GaussianNB and MultinomialNB): Is a simple supervised learning algorithms / statistical classification technique based on Bayes Theorem. Is a fast, accurate and reliable algorithm. Naive Bayes classifiers have high accuracy and are fast on large datasets, because they assume that the effect of a particular feature in a class is independent of other features. The GaussianNB Algorithm uses gaussian distribution and the second one uses the multinomial distribution. The multinomial classification algorithm is the most suitable for word counting. TF-IDF works as well.

Logistic Regression: is a statistical technique that aims to produce, from a set of observations, a model that allows the prediction of values taken by a categorical variable.

Random Forest

K-Nearest Neighbors: KNN algorithm is used to classify by finding the K nearest matches in training data and then using the label of closest matches to predict.

Decision Trees: currently one of the most important supervised learning algorithms. It uses a decision tree to classify the dataset in smaller subsets. The major benefits of decision trees are i) produce intensive results, ii) easy to understand, iii) and well-organized knowledge structure [1].

I have also opted for the following pre-processing techniques:

- Reduction of the subset of data and identification of necessary features.
- Consolidation of product classes into fewer classes for more uniform balancing.
- Text processing to clean complaint narrative.
- Identification of the bag of words using TF-IDF for vectorization of English words.
- Sparse matrix of only unigram scores and bigram scores (I tried Bigrams and Trigrams, but unfortunately there were some problems with the performance of the laptop, leading to the Kernel error).
- Removal of stop words from narratives.

After features identification, the data was divided into train data and test data, so that they could be used to train a classifier and evaluate it, respectively. The first 80% are training sets and the last 20% are test sets. As I defined a shuffle parameter, the data was randomly divided.

The evaluation was through the Cross Validation method, as mentioned before.

5. Benchmark

I applied the different models to the same set of test data and the same set of training data.
The evaluation of the models was based on the metrics presented in the evaluation metrics section.
I will use as a comparison the project developed in a professional scope, for multi-class classification of text collected in surveys in several languages.

Methodology

Based on the literature, work related and projects that have been implemented in the ML area, I will seek to use several tools and libraries such as: Python [10], NumPy [11], Pandas [12], Scikit-learn [13], Sweetviz [14], Streamlit [15] [16], for data exploration, pre-processing and models evaluation. As a final step of this project, a web app was created to interact with my models. To do this I used the export of the model to Pickle and used the Streamline library to develop the app (we can see in more detail in the section dedicated to this theme).

I chose to use a specific jupyter notebook for each phase of the project. Thus:

- *Dataset preparation: 01_data_collection and 02_data_exploration*
- *Feature engineering: 03_data_feature_engineering*
- *Model training: 04_training_models*
- *Improve performance: 05_model_evaluation*
- *Choosing the best model 05_model_evaluation*
- *Deployment model: 06_deployment_model*

1. Data Pre-processing

Here I present all the work done in terms of data pre-processing:

2. Data Cleaning & Text Processing

At this stage, the aim was first to consolidate the various product categories, seeking to reduce them. Then I tried to reduce the dataset by about 90% because it is a very large dataset with extensive complaints narrative text. Already when vectoring the narratives, I cleaned up the text, as explained above.

3. Label-Encoding

Considering the algorithms, I will use, it is important to create a calculated column encoding the product as an integer, because categorical variables are better represented as integers.
The product column in the dataset, contains strings (categorical values), and to prepare them for the extraction of features, I'll want to convert them into numeric values, in a coherent way. In practice I tried to assign a new unique identifier to each different product. Approaches to convert categorical values to numerical:
¶

- *Label-Encoding*: It can only be applied to target variables.
- One-Hot-Encoding: It's applied to training variables.
- Find & Replace: Find categorical values & replace them with a numerical value.

4. Unbalanced Classes

For this project I decided, from the outset, to consolidate the products in order to achieve a better balance of them. Approach:

- *Credit reporting, credit repair services, or other personal consumer reports --> Credit reporting, repair, or other.*
- *Credit reporting -> Credit reporting, repair, or other.*
- *Credit card --> Credit card or prepaid card.*
- *Prepaid card --> Credit card or prepaid card.*
- *Money transfer --> Money transfer, virtual currency, or money service.*
- *Virtual currency --> Money transfer, virtual currency, or money service.*
- *Payday loan, title loan, or personal loan --> Loan.*
- *Student loan --> Loan.*
- *Consumer Loan --> Loan.*
- *Payday loan --> Loan.*
- *Vehicle loan or lease --> Loan.*

5. Bag of Words and Features

To train supervised classifiers, I first transformed the complaint narrative text into a vector of numbers so that the algorithms to be used are able to make predictions (normally several algorithms don't work with raw text). In this case, the TF-IDF like bag of words it will be used to assess the importance of a word for a complaint in a complaints collection (Tables 1):

Table 1 –Example of TfidfVectorizer and CountVectorizer vocabulary.

Each of the 31503 complaints is represented by 10765 features (representing CountVectorizer score of unigrams)
Each of the 31503 complaints is represented by 10765 features (representing TfidfVectorizer score of unigrams)
Each of the 31503 complaints is represented by 10765 features (representing CountVectorizer score of unigrams and clean text)
Each of the 31503 complaints is represented by 10765 features (representing TfidfVectorizer score of unigrams and clean text)
Each of the 31503 complaints is represented by 84278 features (representing CountVectorizer score of bigrams)
Each of the 31503 complaints is represented by 84278 features (representing TfidfVectorizer score of bigrams)
Each of the 31503 complaints is represented by 84278 features (representing CountVectorizer score of bigrams and clean text)
Each of the 31503 complaints is represented by 84278 features (representing TfidfVectorizer score of bigrams and clean text)

6. Stop Words

As the complaint narrative text is long, the removal of stop words is also considered (Figure 10).

Implementation

1. Training and Testing Split Data

Before training and evaluating the models, I opted to split the dataset in test data and training data (80% and 20%). It should return the follow tuples:

- *X_train* and *y_train* - selected training features and their corresponding product labels without text clean.
- *X_test* and *y_test)* - selected training features and their corresponding product labels without text clean.
- *X_test_*clean *and y_test_clean* - selected training features and their corresponding product labels with complaints narrative text clean.

2. Training and Evaluation Model

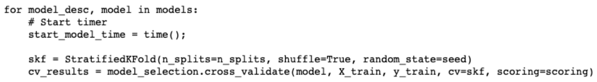
The classification process followed the following approach:

1. data loading.
2. cleaning of the narratives text.
3. division of the dataset into training and testing.
4. creation of the bag of words using TF-IDF for the text of the narratives.
5. train the model and calculate the accuracy, F-1 score, recall e prediction (Figure 10).
6. Finally, a classification report was generated by each model (the two with the best performance - Figure 11).

I simulated this process eight times for each test definition.

3. Model Performance Metrics

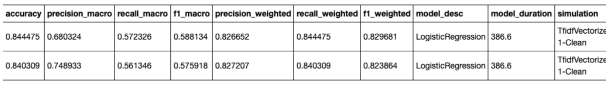
To estimate the accuracy and the other metrics I used the cross-validation strategy, using the *StratifiedKFold* method (explained above) which split the dataset in four parts for training (80%) and one for testing (20%) (Figure 10).



```
for model_desc, model in models:
    # Start timer
    start_model_time = time();

    skf = StratifiedKFold(n_splits=n_splits, shuffle=True, random_state=seed)
    cv_results = model_selection.cross_validate(model, X_train, y_train, cv=skf, scoring=scoring)
```

Figure 10 – Part of code to train and evaluate the models.



accuracy	precision_macro	recall_macro	f1_macro	precision_weighted	recall_weighted	f1_weighted	model_desc	model_duration	simulation
0.844475	0.680324	0.572326	0.588134	0.826652	0.844475	0.829681	LogisticRegression	386.6	TfidfVectorize 1-Clean
0.840309	0.748933	0.561346	0.575918	0.827207	0.840309	0.823864	LogisticRegression	386.6	TfidfVectorize 1-Clean

Figure 11 – Example of the model training output (one line for each K and each model).

4. Challenges

As I have already mentioned, the most important thing would be to be able to complete this project, despite the technical difficulties that have occurred related to the performance of my local laptop. The volume of data was large, the identified bag of words also had a considerable volume, which caused the kernel to be constantly interrupted in the training of some models. Therefore, I chose not to use the bigrams TF-IDF (which I believe will give more precision to the model), not to train some models and not to make big parameter changes, so that in the end it would be possible to have one or two models trained and with an accuracy above 80% that I consider good to follow the project.

5. Refinement

Due to the performance challenge of my machine, during the implementation of the models several tests and simulations were carried out, leading to the refinement of the models according to the needs.
New models were defined to train the data set, variations of the models already selected with the use of different parameters and some models were also removed due to performance problems in their execution:

Unbalanced classes: I have chosen to consolidate the products in other products classes.

Bag of words: To create the bag of words, I considered TfidfVectorizer and CountVectorizer. Some models performed better than others, but in the end, I chose the simulations that used the TfidfVectorizer. In this case, TFIDF will be used to assess the importance of a word for a complaint in a complaints collection.

Cleaning of the complaint's narratives text: I considered carrying out the cleaning of the narratives to apply the TfidfVectorizer and the CountVectorizer. Following this, eight pre-processed data subsets were created to test the model's performance.

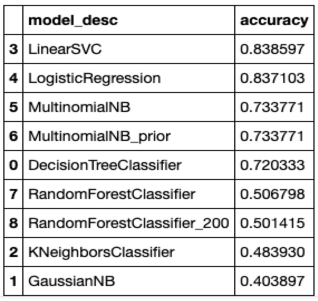
Parameters: A number of parameters were set as the simulations were carried out. I used cross validate to analyze the accuracy and the remaining macro metrics (in weighted and average terms).
In the end, due to performance problems, I chose not to apply several changes in the parameters, in order to have a good classifier that could be used in the evaluation phase.

Results

1. Model Evaluation and Validation

After training the various models, and the various simulations, I validated the results using the generated output. The table below represents the global average of accuracy per trained models.

Table 2 –Accuracy across model.

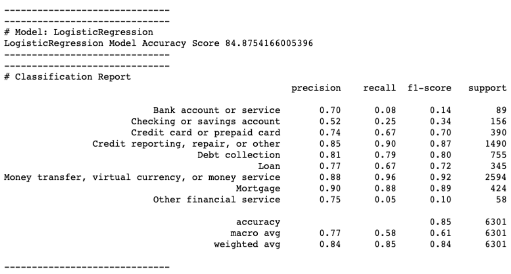


	model_desc	accuracy
3	LinearSVC	0.838597
4	LogisticRegression	0.837103
5	MultinomialNB	0.733771
6	MultinomialNB_prior	0.733771
0	DecisionTreeClassifier	0.720333
7	RandomForestClassifier	0.506798
8	RandomForestClassifier_200	0.501415
2	KNeighborsClassifier	0.483930
1	GaussianNB	0.403897

I verified that the two best models, by large margin, are the Logistic Regression models with 84% and the LinearSVC model also with 84%.

I selected these two models in specific to evaluate the performance of the predictions, for that I generated a classification report with the metrics that I propose to validate, as well as the confusion matrix, where it is easy to perceive the differences between the actual products and the predicted products.

Logistic Regression report



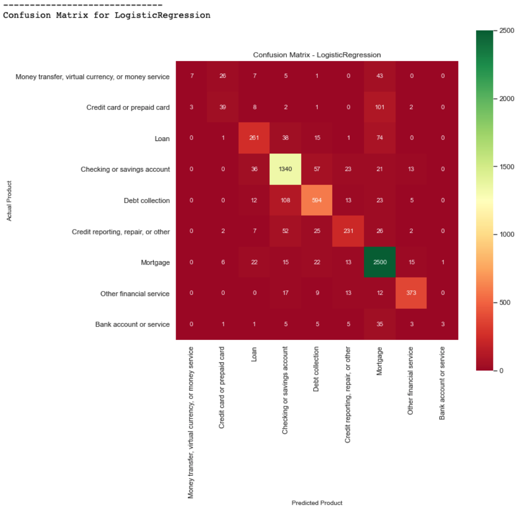
```
------------------------------
------------------------------
# Model: LogisticRegression
LogisticRegression Model Accuracy Score 84.8754166005396
------------------------------
------------------------------
# Classification Report
                                                 precision    recall  f1-score   support

                         Bank account or service       0.70      0.08      0.14        89
                     Checking or savings account       0.52      0.25      0.34       156
                     Credit card or prepaid card       0.74      0.67      0.70       390
              Credit reporting, repair, or other       0.85      0.90      0.87      1490
                                 Debt collection       0.81      0.79      0.80       755
                                            Loan       0.77      0.67      0.72       345
Money transfer, virtual currency, or money service       0.88      0.96      0.92      2594
                                        Mortgage       0.90      0.88      0.89       424
                         Other financial service       0.75      0.05      0.10        58

                                        accuracy                           0.85      6301
                                       macro avg       0.77      0.58      0.61      6301
                                    weighted avg       0.84      0.85      0.84      6301

------------------------------
```

Figure 12 – Classification report for Logistic Regression model.

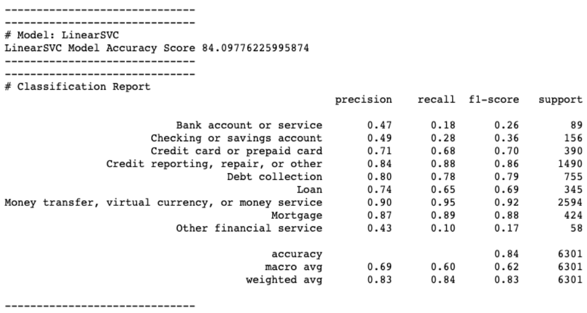


```
------------------------------
Confusion Matrix for LogisticRegression
```

	Money transfer, virtual currency, or money service	Credit card or prepaid card	Loan	Checking or savings account	Debt collection	Credit reporting, repair, or other	Mortgage	Other financial service	Bank account or service
Money transfer, virtual currency, or money service	7	26	7	5	1	0	43	0	0
Credit card or prepaid card	3	39	8	2	1	0	101	2	0
Loan	0	1	261	38	15	1	74	0	0
Checking or savings account	0	0	36	1340	57	23	21	13	0
Debt collection	0	0	12	108	594	13	23	5	0
Credit reporting, repair, or other	0	2	7	52	25	231	26	2	0
Mortgage	0	6	22	15	22	13	2500	15	1
Other financial service	0	0	0	17	9	13	12	373	0
Bank account or service	0	1	1	5	5	5	35	3	3

Figure 13 – Confusion matrix for Logistic Regression model.

Linear SVC report



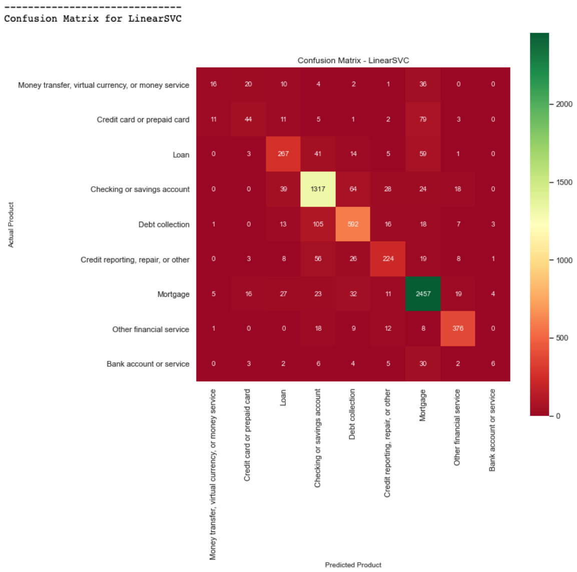
```
------------------------------
------------------------------
# Model: LinearSVC
LinearSVC Model Accuracy Score 84.09776225995874
------------------------------
------------------------------
# Classification Report
                                                 precision    recall  f1-score   support

                         Bank account or service       0.47      0.18      0.26        89
                     Checking or savings account       0.49      0.28      0.36       156
                     Credit card or prepaid card       0.71      0.68      0.70       390
              Credit reporting, repair, or other       0.84      0.88      0.86      1490
                                 Debt collection       0.80      0.78      0.79       755
                                            Loan       0.74      0.65      0.69       345
Money transfer, virtual currency, or money service       0.90      0.95      0.92      2594
                                        Mortgage       0.87      0.89      0.88       424
                         Other financial service       0.43      0.10      0.17        58

                                        accuracy                           0.84      6301
                                       macro avg       0.69      0.60      0.62      6301
                                    weighted avg       0.83      0.84      0.83      6301

------------------------------
```

Figure 14 – Classification report for Linear SVC model.

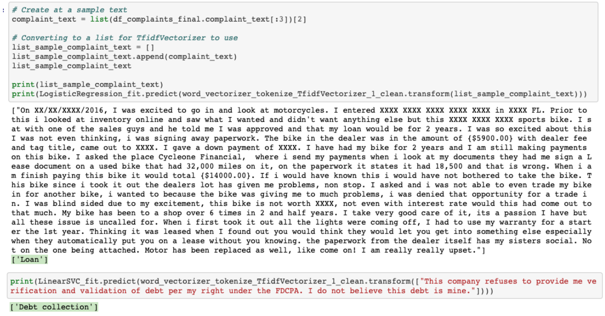


```
------------------------------
Confusion Matrix for LinearSVC
```

	Money transfer, virtual currency, or money service	Credit card or prepaid card	Loan	Checking or savings account	Debt collection	Credit reporting, repair, or other	Mortgage	Other financial service	Bank account or service
Money transfer, virtual currency, or money service	16	20	10	4	2	1	36	0	0
Credit card or prepaid card	11	44	11	5	1	2	79	3	0
Loan	0	3	267	41	14	5	59	1	0
Checking or savings account	0	0	39	1317	64	28	24	18	0
Debt collection	1	0	13	105	592	16	18	7	3
Credit reporting, repair, or other	0	3	8	56	26	224	19	8	1
Mortgage	5	16	27	23	32	11	2457	19	4
Other financial service	1	0	0	18	9	12	8	376	0
Bank account or service	0	3	2	6	4	5	30	2	6

Figure 15 – Confusion matrix for Linear SVC model.

It is noticeable that the model performs well (example in Figure 16), but due to the limitations (unigrams, no removal of outliers, and setting of new parameters) it is also noticeable that when the narratives do not address financial issues, the model randomly assigns a product.



```
# Create at a sample text
complaint_text = list(df_complaints_final.complaint_text[:3])[2]

# Converting to a list for TfidfVectorizer to use
list_sample_complaint_text = []
list_sample_complaint_text.append(complaint_text)
list_sample_complaint_text

print(list_sample_complaint_text)
print(LogisticRegression_fit.predict(word_vectorizer_tokenize_TfidfVectorizer_1_clean.transform(list_sample_complaint_text)))
```

```
["On XX/XX/XXXX/2016, I was excited to go in and look at motorcycles. I entered XXXX XXXX XXXX XXXX XXXX in XXXX FL. Prior to
this i looked at inventory online and saw what I wanted and didn't want anything else but this XXXX XXXX XXXX sports bike. I s
at with one of the sales guys and he told me I was approved and that my loan would be for 2 years. I was so excited about this
I was not even thinking, i was signing away paperwork. The bike in the dealer was in the amount of {$5900.00} with dealer fee
and tag title, came out to XXXX. I gave a down payment of XXXX. I have had my bike for 2 years and I am still making payments
on this bike. I asked the place Cycleone Financial,  where i send my payments when i look at my documents they had me sign a L
ease document on a used bike that had 32,000 miles on it, on the paperwork it states it had 18,500 and that is wrong. When i a
m finish paying this bike it would total {$14000.00}. If i would have known this i would have not bothered to take the bike. T
his bike since i took it out the dealers lot has given me problems, non stop. I asked and i was not able to even trade my bike
in for another bike, i wanted to because the bike was giving me to much problems, i was denied that opportunity for a trade i
n. I was blind sided due to my excitement, this bike is not worth XXXX, not even with interest rate would this had come out to
that much. My bike has been to a shop over 6 times in 2 and half years. I take very good care of it, its a passion I have but
all these issue is uncalled for. When i first took it out all the lights were coming off, I had to use my warranty for a start
er the 1st year. Thinking it was leased when I found out you would think they would let you get into something else especially
when they automatically put you on a lease without you knowing. the paperwork from the dealer itself has my sisters social. No
t on the one being attached. Motor has been replaced as well, like come on! I am really really upset."]
['Loan']
```

```
print(LinearSVC_fit.predict(word_vectorizer_tokenize_TfidfVectorizer_1_clean.transform(["This company refuses to provide me ve
rification and validation of debt per my right under the FDCPA. I do not believe this debt is mine."])))
```

```
['Debt collection']
```

Figure 16 – Predicting using Logistic Regression model.

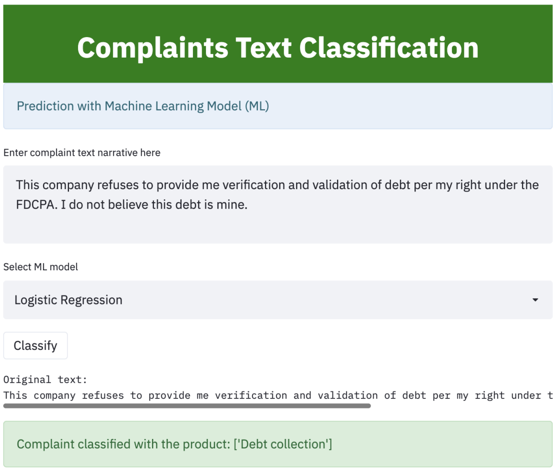
2. Justification

Many projects can be found in literature or similar under the multi-class text classification [29], [30], [31] and [32]. For problems of this kind or related, for similar data sets, the models used are usually Logistic Regression and Linear SVC, where the accuracy is always between 70% and 85%, depending on the use of more sophisticated techniques for data pre-processing and configuration of best parameters.

In my particular case, compared to the other models (using logistic regression), which I implemented in real situations, for a multi-class problem, for the classification of text collected through surveys, in several languages (languages including Russian, Bulgarian, Romanian, Portuguese, English, Polish, Ukrainian and Kazakhstan) with an accuracy of 75%, I can assume that during the evaluation of the metrics, the two models defined here have a better performance.

3. Deployment Model

As proposed, a web application was created to interact with the model (Figure 17).



Complaints Text Classification

Prediction with Machine Learning Model (ML)

Enter complaint text narrative here

This company refuses to provide me verification and validation of debt per my right under the FDCPA. I do not believe this debt is mine.

Select ML model

Logistic Regression

Classify

```
Original text:
This company refuses to provide me verification and validation of debt per my right under t
```

Complaint classified with the product: ['Debt collection']

Figure 17 Example of web app to predict best product for each complaint narrative text by model.

Conclusions

The purpose of this project was to find the best model for classifying consumer finance complaints into products.

With this work I explored several ML techniques, trained and tested several classification models and validated their precision and performance.
In the end I chose two models, with the best performance for validation.

I identified that logistic regression and supporting vector machines are the models with the highest prediction accuracy on the final data set.

These models have been used extensively and, through the sample, it is concluded that they can make a considerably good prediction, despite some flaws.

After some tests it turns out that some of the complaints are poorly classified because they address more than one particular issue. Or one can also verify that new complaints that have nothing to do with the financial subject are classified to products that have nothing to do with it. Regarding the first problem, these are errors that will normally happen (a complaint can say debit card and loans and is classified for one of them). Regarding the second problem, I can advance that in future work a product can be created ("Others") where these complaints can "enter".

Various simulations were carried out to evaluate the models, from the use of tokenizers., only n grams or bigrams and cleaning up or not the narratives of the complaints. In total eight refinements were made to validate the models. I maintained the use of default parameters for performance reasons.

In this context, bag of words techniques was used to tokenize the words of the complaints, TF-IDF to translate the frequency of words, stop words to remove common terms, clean and reduce data set and product consolidation to test and split the complaints.

The challenge in this project were the technical difficulties with the performance of my local laptop for the initial data set and training algorithms.

Future Work

In terms of research and future work, it is possible to greatly improve this classification model by concentrating on different methods of data manipulation and the combination of products. In addition, more balanced data and larger samples could be useful for forecasting, because due to technical difficulties, I was not able to test with Bigrams, with more complex models, and with a considerably large data set. It can also be tried to remove outliers from narratives to ensure that the model is not muddled.

References & Bibliography

1. Iqbal, Muhammad & Yan, Zhu. (2015). SUPERVISED MACHINE LEARNING APPROACHES: A SURVEY. International Journal of Soft Computing. 5. 946-952. 10.21917/ijsc.2015.0133.
2. Monkeylearn.com. Accessed 14 Nov. 2020. Available at: https://monkeylearn.com/blog
3. New Features in the CFPB Consumer Complaint Database. Accessed 14 Nov. 2020. Available at: https://www.jdsupra.com/legalnews/new-features-in-the-cfpb-consumer-40931/
4. cfpb.github.io. Accessed 14 Nov. 2020. Available at: https://cfpb.github.io/api/ccdb/
5. paperswithcode.com/task/text-classification Accessed 14 Nov. 2020. Available at: https://paperswithcode.com/task/text-classification
6. kaggle - Consumer Complaint Database. Accessed 14 Nov. 2020. Available at: https://www.kaggle.com/selener/consumer-complaint-database
7. Catalog.data.gov - Consumer Complaint Database. Accessed 14 Nov. 2020. Available at:

https://catalog.data.gov/dataset/consumer-complaint-database

8. 2014 intro supervised learning. Accessed 14 Nov. 2020. Available at: https://sebastianraschka.com/Articles/2014_intro_supervised_learning.html
9. IT support ticket classification using machine learning and ml model deployment. Accessed 14 Nov. 2020. Available at: https://medium.com/@karthikkumar_57917/it-support-ticket-classification-using-machine-learning-and-ml-model-deployment-ba694c01e416
10. Python. Accessed 14 Nov. 2020. Available at: https://www.python.org/
11. Numpy. Accessed 14 Nov. 2020. Available at: https://numpy.org/
12. Pandas.pydata. Accessed 14 Nov. 2020. Available at: https://pandas.pydata.org/
13. Scikit-learn. Accessed 14 Nov. 2020. Available at: https://scikit-learn.org/stable/
14. Fast and insightful eda exploratory data analysis using sweetviz 1. Accessed 14 Nov. 2020. Available at: https://towardsdatascience.com/fast-and-insightful-eda-exploratory-data-analysis-using-sweetviz-1-1-3dd0f2389b6c
15. Streamlit. Accessed 14 Nov. 2020. Available at: https://www.streamlit.io/
16. Build-and-deploy-machine-learning-web-app-using-pycaret-and-streamlit. Accessed 14 Nov. 2020. Available at: https://towardsdatascience.com/build-and-deploy-machine-learning-web-app-using-pycaret-and-streamlit-28883a569104
17. Tableau. Accessed 14 Nov. 2020. Available at: https://www.tableau.com/
18. Build and deploy your first machine learning web app. Accessed 14 Nov. 2020. Available at: https://www.kdnuggets.com/2020/05/build-deploy-machine-learning-web-app.html
19. Sklearn Metrics Accuracy Score. Accessed 14 Nov. 2020. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html#sklearn.metrics.accuracy_score
20. Predictive modeling, supervised machine learning, and pattern classification. Accessed 14 Nov. 2020. Available at: https://sebastianraschka.com/Articles/2014_intro_supervised_learning.html#machine-learning-and-pattern-classification
21. CRISP-DM, SEMMA e KDD: conheça as melhores técnicas para exploração de dados. Accessed 14 Nov. 2020. Available at: https://paulovasconcellos.com.br/crisp-dm-semma-e-kdd-conhe%C3%A7a-as-melhores-t%C3%A9cnicas-para-explora%C3%A7%C3%A3o-de-dados-560d294547d2
22. A Comprehensive Guide to Understand and Implement Text Classification in Python. Accessed 14 Nov. 2020. Available at: https://www.analyticsvidhya.com/blog/2018/04/a-comprehensive-guide-to-understand-and-implement-text-classification-in-python
23. Organizing machine learning projects: project management guidelines. Accessed 14 Nov. 2020. Available at: https://www.jeremyjordan.me/ml-projects-guide/#data
24. Report on Text Classification using CNN, RNN & HAN. Accessed 14 Nov. 2020. Available at: https://medium.com/jatana/report-on-text-classification-using-cnn-rnn-han-f0e887214d5f
25. Feature Selection with sklearn and Pandas. Accessed 14 Nov. 2020. Available at: https://towardsdatascience.com/feature-selection-with-pandas-e3690ad8504b
26. Hyperparameter tuning for machine learning models. Accessed 14 Nov. 2020. Available at: https://www.jeremyjordan.me/hyperparameter-tuning
27. analyzing-text-classification-techniques-on-youtube-data . Accessed 14 Nov. 2020 https://www.codementor.io/@rohitagrawalofficialmail/analyzing-text-classification-techniques-on-youtube-data-x5sa1cdvw
28. Understanding-stratified-cross-validation. Accessed 14 Nov. 2020. https://stats.stackexchange.com/questions/49540/understanding-stratified-cross-validation
29. Multi-Class Text Classification with Scikit-Learn Accessed 14 Nov. 2020. https://towardsdatascience.com/multi-class-text-classification-with-scikit-learn-12f1e60e0a9f
30. Text Classification Analysis. Accessed 14 Nov. 2020. https://github.com/agrawal-rohit/Text-Classification-Analysis
31. Multi-class text classification (TFIDF) - Accessed 14 Nov. 2020. https://www.kaggle.com/selener/multi-class-text-classification-tfidf
32. Project_4_Consumer_Complaint_Text_MultiClassifier - Accessed 14. https://github.com/xianjinseow92/Data-Science-Projects/tree/master/Project_4_Consumer_Complaint_Text_MultiClassifier
33. Fast and insightful Exploratory Data Analysis using Sweetviz 1.1. Accessed 14 Nov. 2020. https://towardsdatascience.com/fast-and-insightful-eda-exploratory-data-analysis-using-sweetviz-1-1-3dd0f2389b6c
34. How can I explain my ML models to the business? Accessed 14 Nov. 2020. https://towardsdatascience.com/how-can-i-explain-my-ml-models-to-the-business-dc4d97997d64
35. Multi-Class Metrics Made Simple, Part I: Precision and Recall - Accessed 14 Nov. 2020. https://towardsdatascience.com/multi-class-metrics-made-simple-part-i-precision-and-recall-9250280bddc2
36. unigrams & bigrams (tf-idf) less accurate than just unigrams (ff-idf)? - Accessed 14 Nov. 2020. https://stackoverflow.com/questions/12247768/unigrams-bigrams-tf-idf-less-accurate-than-just-unigrams-ff-idf
37. How to train_test_split : KFold vs StratifiedKFold - Accessed 14 Nov. 2020. https://towardsdatascience.com/how-to-train-test-split-kfold-vs-stratifiedkfold-281767b93869
38. Deploy a Machine Learning Model using Streamlit Library - Accessed 14 Nov. 2020. https://www.geeksforgeeks.org/deploy-a-machine-learning-model-using-streamlit-library/?ref=rp
39. Humanizing Customer Complaints using NLP Algorithms - Accessed 14 Nov. 2020. https://towardsdatascience.com/https-medium-com-vishalmorde-humanizing-customer-complaints-using-nlp-algorithms-64a820cef373
40. Feature Selection Techniques in Machine Learning with Python - Accessed 14 Nov. 2020. https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e
41. Feature Selection Techniques in Machine Learning with Python - Accessed 14 Nov. 2020. https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e

Complaints Text Classification

UDACITY MACHINE LEARNING ENGINEER NANODEGREE

Sérgio da Costa

November 14, 2020