**Capstone Project Documentation**

# Process overview

The following diagram shows the overall end-to-end process for defining, designing, and delivering as a business solution.

Diagram

Description automatically generated

# Problem statement

In the modern Internet-driven world customers have a lot of power to spread the word about the companies they get services from. A positive customer experience will be greatly beneficial for both to retain and gain customers as the experience is shared by them, a negative customer experience can also impact negatively to a company.

It is estimated that over 90% customers do not complain about a negative experience. This can be valuable insight to provide a better service and to retain customers. It is not only important to provide a solution, but it is essential to provide a solution to these complaints as fast as possible as traditional ways of communication are on the decline.

Furthermore, there’s a constant challenge from other companies that invest in modern technologies to provide a better service which pushes all other companies to use the latest and the most effective technologies which leads us to find a solution using Machine Learning.

Over the past couple of decades, a lot of companies have had success using chatbots and analysis of text data to address this. A comprehensive solution that uses latest Machine Learning and Natural Language Processing technologies has a great potential.

# Industry / Domain

The industry that this project focuses on is the Finance sector. This industry covers domains ranging from investments, banking, credit management to modern fintech companies. The competition between the companies is at the highest it has ever been (ex- banks vs fintech) Finance industry also has a very reactive customer base as it is handling people’s finances. The customer engagement is high. Although the domain of this project is focused on this sector, the methodologies can be used as it is in most other industries that have a similar B2C business process.

# Stakeholders

The stakeholders of this project will be the management of finance companies. This project attempts to provide a solution as a generic solution that can be applied by any finance company rather than focusing on a single company. To address this, the dataset used in the project takes data from over 2000+ different finance companies. Through this project I will be analysing how companies can take from one day to upto 10 days on average currently to provide a resolution to a customer complaint. These differences can have a significant impact on customer experience. The objective is to bring attention of the stakeholders to these issues and provide a solution through technology that in return will benefit the company.

# Business question

The business question I am answering in this project is “How can we use Customer Complaints Classification to improve customer service?” I’m answering this business question as an improvement to overall customer experience which will in return increase company’s revenue. The exact business value of answering this question would differ from company to company. The project aims at getting upto 80% accuracy and the implication of false positives and negatives in this domain is moderate.

# Data question

The data question answered in this project is predicting the text label of the customer complaints type as a multi-class classification by analysing a dataset of text complaints by customers.

# Data

The data was sourced through Kaggle, which is also publicly available on the US Consumer Financial Protection Bureau website making it a reliable source. The Consumer Complaint Database is a collection of complaints about consumer financial products and services that we sent to companies for response. This dataset contains data from over 2000 companies from the US. It is updated by the US Consumer Financial Protection Bureau and data are released 15 days after the company responds.

The following chart shows the top 10 companies with complaints in the dataset used.

Chart, bar chart

Description automatically generated

The dataset used is about 170MB in size with 556,000 rows in total. For the scope of this project of customer complaints analysis, the dataset only has about 65,000 rows with customer complaints and the rest were not usable for the Natural Language Processing (NLP) and modelling. However, the remaining features still provided some insights about the dataset. The dataset contains the following features and for modelling, only the Product & Consumer Complaint Narrative was used.

* Product (Complaint Type - Predicting Label)
* Sub Product
* Issue
* Sub Issue
* Consumer Complaint Narrative (Text complaint used for NLP analysis and modelling)
* Company Public Response
* Company
* State
* Zip Code
* Tags
* Consumer Consent Provided
* Submitted Via
* Date Sent to Company
* Company Response to Consumer
* Timely Response
* Consumer Disputed
* Complaint Id

# Data science process

## Data analysis

The Exploratory Data Analysis (EDA) process was firstly done on the main predicting label (Complaint Type) and after some cleaning it was cleaned and narrowed down to 6 main types as follows, which was a balanced dataset.

Chart, pie chart

Description automatically generated

One of the key highlights was the “Submitted Via” feature which had 6 different values yet only the “Web” type (online complaints) had a text complaint useful for modelling purposes. And the “Web” type also had about 80% of blank data, but we still had about 65000 rows for analysis after cleaning.

Chart, bar chart

Description automatically generated

Even though the other Submitted Types did not contain a text complaint for analysis, the data available in other columns were still very useful for understanding the business problem. The following chart shows how different types takes different times for responses, which was one of the key focus areas in the project.

Chart, bar chart

Description automatically generated

## Modelling

The 2 main features used for this project were the **Product** (Complaint Type) and the **Consumer Complaint Narrative.** The issue, sub product & sub issue had a correlation to the complaint and were incomplete for meaningful analysis.

The text cleaning was mainly done through Regular Expressions and NLP cleaning techniques such as removal of punctuations, Stop Words and then Lemmatized. 3 Vectorization methods were used and compared for each model used (Count, TF-IDF, N-Gram – 2 & 3 word N-Gram).

Modelling was done with the following modelling techniques.

* + Multinomial Naïve Bayes
  + Support Vector Machine
  + Random Forest
  + Gradient Boosting
  + XG Boost
  + LSTM - Recurrent Neural Network
  + 1D Convolutional Neural Network

In addition, the following NLP modelling techniques and steps were used

* Topic Modelling (Latent Dirichlet Allocation - LDA)
* Sentiment Analysis (Textblob)
* Emotion Detection (text2emotion)
* Text Summarization (Sumy)
* Additional Modelling on Company Response Text Messages using the best performing model.

All development and training were done on Google Colab platform with the utilization of a GPU. The two Neural Network models took the longest, yet it was still in minutes under half an hour.

The model was selected with the focus on Accuracy and F1 scores on each label category. **XGBoost** was the best performing model.

The accuracy scores on test data:

A picture containing text

Description automatically generated

Table

Description automatically generated

## Outcomes

Following are the key findings as a summary.

* We can predict customer complaints category with an accuracy up to 86%
* XGBoost was the best performing model.
* Using Term-Frequency at a single word level yields the best results when vectorizing.
* Additional feature columns used did not contain enough data for further meaningful modelling in this dataset, but it helped understand the problem better.
* Text Summarization helps as an additional step when we have longer texts to summarize the problem. But this does not work well on shorter texts less than 2 sentences.
* Emotion detection on text does not return meaningful results on complaints data
* Topic modelling can also be used to categorize when labelled data are unavailable.
* Sentiment analysis helps distinguish between customer dissatisfaction and a formal complaint.

# Data & Business Answer

The data answer was successfully answered with a high accuracy over the expected 80% mark on all complaint categories. The satisfactory answering of the data question confirmed that we can use Customer Complaints Classification to improve customer service by predicting labels correctly. The goal is to use the predicted labels in a system where we direct customers to the right person / department to provide them a quick and accurate solution.

# Response to stakeholders

The high accuracy level, the short amount of time required for training and the need of reasonable amount of resources to build and train the final model gives us confidence to deploy the model and implement this as an end to end solution within a financial company. This also encourages more customers to submit their complaints via an online system making the process more efficient and easier to manage.

# End-to-end solution & Implementation

The following diagram shows how we can combine different inputs of complaint data, extracting the text, (image to text is out of scope for this project and only used as a demo in the diagram) and bring it together to connect with our deployed model. A real time request to trained data will be sent and the model will provide a classification and a summarization that would direct the customer to the relevant service.

Diagram

Description automatically generated

# References

* Kaggle <https://www.kaggle.com/kaggle/us-consumer-finance-complaints>
* US Consumer Financial Protection Bureau <https://www.consumerfinance.gov/data-research/consumer-complaints/>
* [Hands-On Machine Learning with Scikit-Learn and TensorFlow Book by](https://dl.acm.org/doi/book/10.5555/3153997)[[Aurelien Geron](https://dl.acm.org/doi/book/10.5555/3153997)](https://www.amazon.com.au/s/ref=dp_byline_sr_book_1?ie=UTF8&field-author=Aurelien+Geron&text=Aurelien+Geron&sort=relevancerank&search-alias=books-single-index)[(Author)](https://dl.acm.org/doi/book/10.5555/3153997)
* TensorFlow Developer Certificate in 2022: Zero to Mastery – Online Course Contents <https://www.udemy.com/course/tensorflow-developer-certificate-machine-learning-zero-to-mastery/>
* Data Science from Scratch: First Principles with Python Book by Joel Grus
* <https://towardsdatascience.com/text2emotion-python-package-to-detect-emotions-from-textual-data-b2e7b7ce1153>
* <https://medium.com/batteries-included/sentiment-analysis-using-lstm-models-nlp-part-3-6733e0aeb334>
* <https://medium.com/@ondenyi.eric/extractive-text-summarization-techniques-with-sumy-3d3b127a0a32>
* Datacamp resources on Classification models.
* Institute of Data course materials.