# Class Project 2 – Predictive Modeling

The Consumer Financial Protection Bureau (CFPB) is a U.S. government agency that makes sure financial companies treat their customers fairly. Their website allows customers of financial services to file complaints against financial companies and banks against unfair treatment if these companies are unable to resolve complaints to the customer’s satisfaction.

When customers choose to complain to the CFPB, financial companies incur additional costs to resolve such complaints.

On receipt, the CFPB routes complaints to the financial companies, who generally respond to the consumer within 15 days. Once a response is provided, one of two things can happen:

1. In most cases, consumers accept the response or remediation offered by the financial companies,
2. In other cases, they choose to dispute the resolution offered by the company. (flagged in the 'Consumer disputed?' field). In these situations, the bank has to perform additional investigations, and possibly offer further relief to the customers. As a result, the cost of dealing with disputes can be high.

The original dataset for this project has over 2 million anonymized recent records, and covers 6000+ financial providers of all varieties. It can be downloaded following the instructions at <https://www.consumerfinance.gov/data-research/consumer-complaints/>. The website also provides additional information on the data, including the data dictionary.

For this project, we will use only the data till 2017, and only for the top 5 banks in the US. In order to make sure we are all working off the same data, **we will use the file *complaints\_25Nov21.csv* available in Jupyterhub under the shared/ folder.**

The cost structure:

1. On average, it costs the banks $100 to resolve, respond to and close a complaint that is not disputed.
2. On the other hand, it costs banks an extra $500 to resolve a complaint if it has been disputed. (This $500 is on top of the $100 they have already spent.)
3. Extra diligence: If the banks know in advance which complaints will be disputed, they can perform “extra diligence” during the first round of addressing the complaint with a view to avoiding eventual disputes. Performing extra diligence costs $90 per complaint, and provides a guarantee that the customer will not dispute the complaint. But performing the extra diligence is wasted money if the customer would not have disputed the complaint.

You are required to create a model that can help the banks identify complaints that will end in a dispute. The goal is to minimize total financial costs, and if the banks can identify future disputes they can avoid the larger costs by performing the cheaper extra diligence in advance.

Hint: Think about Calculating Total Cost in Dollars

* The moment a complaint enters the CFPB’s system, there is $100 cost to resolve it. This applies to every complaint.
* After that, if a complaint’s resolution is disputed by the customer, an additional $500 has to be spent (for a total cost for such cases to be $600).
* But the bank can intervene in advance by spending an extra $90 for extra diligence, and that can make sure the complaint’s resolution is not disputed.

While we can’t prevent complaints from coming to the CFPB, we can reduce total costs by identifying the complaints that are likely to be disputed, and doing the extra due diligence for them. This extra due diligence will cost us an extra $90 per complaint, but save us the additional $500 to resolve the complaint after the dispute. But obviously, the bank would not want to spend this extra money on complaints that would not have been disputed anyway.

Your task is to create a predictive model that can help the banks keep their total complaint related costs low.

Follow the instructions below and answer the multiple choice questions that follow.

1. Explore the data, familiarize yourself with the fields and perform some EDA.
2. Set your X (predictor) and y (predicted) variables.
   1. Use only the below variables as your predictors. Ignore the other variables in the dataset.

'Product', 'Sub-product', 'Issue', 'State', 'Tags', 'Submitted via', 'Company response to consumer', 'Timely response?'

* 1. Use 'Consumer disputed?' as your y-variable. Be sure to convert your y-variable to 0s and 1s so your model can use it.

For example, you can use label encoder as below, or any other method you are comfortable using:

**from** sklearn **import** preprocessing

le **=** preprocessing**.**LabelEncoder**()**

y **=** le**.**fit\_transform**(**complaints**[**'Consumer disputed?'**])**

1. Split your data into a test and train set. Use an 80/20 train-test split, and random\_state=123 for the train-test split.

For example, using the below, appropriately adjusted to the variable names you are using:

X\_train**,** X\_test**,** y\_train**,** y\_test **=** train\_test\_split**(**X**,** y**,** test\_size**=**0.2**,** random\_state **=** 123**)**

1. Check what proportion of complaints in your training dataset are disputed. If this proportion is less than 30%, use random undersampling with random\_state = 123 to balance your dataset.

For example, you could use the below (adjusted for your choice of variables etc)

**from** imblearn**.**under\_sampling **import** RandomUnderSampler

undersampler **=** RandomUnderSampler**(**random\_state**=**123**)**

X\_train**,** y\_train **=** undersampler**.**fit\_resample**(**X**,** y**)**

1. Train a predictive model to predict whether a complaint would be disputed using XGBoost Classifier using random\_state=123

For example, using the below:

model\_xgb **=** XGBClassifier**(**random\_state **=** 123**)**

1. Evaluate the model on the test set, and create the classification report and confusion matrix. (Remember, when we say ‘True Positive’, ‘False Negative’ etc, the second word, positive or negative, denotes the ground truth; and the first word, True or False, indicates whether we predicted correctly.)
2. Calculate the total cost in dollars for the test set. Establish the ‘base-case’, ie the total cost if you were not using a model, using the test set only.

Use the cost structure explained earlier (ie, $600 total for every disputed complaint, and $100 for every non-disputed complaint, and $90 for the extra due diligence.)

1. Now calculate the total cost in dollars based on the model results in the confusion matrix. The below graphic might help you. But you are free to use your own methods.

A chart of a box

Description automatically generated with medium confidence

1. The cost in the default model is not the lowest cost. Change the classification threshold on the model to calculate the lowest total cost you can achieve.