We also investigated a topic modeling approach to identify several key areas Delta Faucet could improve in their maintenance pipeline, and we found that there is a significant difference in the performance of different sampling techniques on our dataset.

The methods chosen for the project were Logistic Regression and Random Forest Classifier. Logistic Regression predicts discrete values (binary values 0/1, true/false, yes/no), given a set of independent variables, and Random Forest represents a group of decision trees.

In our experiments, we discovered that when using a combination of sampling methods, the Logistic Regression model has a tendency to overfit. This causes lower performances in precision, accuracy, and recall.

Using a GridSearchCV pipeline, we found that there is no statistically significant difference between using different sampling techniques on our dataset. However, we have also experimented with using a minimally preprocessed dataset (i.e., no customized word removal and spell checking), but there does not seem to have any performance changes as well.

We found that we had many more samples of class "5 – Very Satisfied" than the others. This indicates that we are dealing with a dataset with class imbalance.

We used the Python library 'imblearn', which implemented 3 potential fixes to this issue:

1. **Fix 1: Oversampling** methods copy examples in the minority class (all other than class 5 in our case) or synthesize new examples from examples in the minority class. We primarily used Borderline SMOTE oversampling.

2. **Fix 2: Under-sampling** methods remove/select a subset of examples from the majority class (class 5 in our case). We primarily used random under-sampling.

3. **Fix 3:** We also investigated a combination of oversampling and under-sampling techniques.

See the figures to the side to see the change of class sizes with the different fixes.

To improve our prediction accuracy from baseline, we wish to focus on four main areas:

- Investigating other better suited columns for solving the business problem.
- A shift of focus from TF-IDF to Word2Vec to see if changing how the words of a dataset are numerically represented would change how a model behaves.
- **Topic Modelling** is a technique that helps produce "topics" of words that you would expect to occur often together in your dataset. Doing so will help us analyze the most frequently occurring words in the customer satisfaction survey.

- **Research on other types of models**, e.g., other machine learning models or deep learning models.

**ACKNOWLEDGEMENTS:** To the DPC and the DataMine team for your encouragement! Our team’s Purdue Data Mine Senior Data Scientist advisor, *David Glass* and Delta Faucet Corporate Partner Mentors, *Nathan Johns* and *Neha Kichambare*, for sharing countless resources with us along our journey.