Predicting Maintenance Case Difficulty with Text Reviews using NLP

Current Team Members: Min Lu, Sophie Kwon, Vandana Prabhu, Madhuri Vempati, Varun Karlekar
Previous Members: Aayu Neema, Alice Chan, Bibire Falodun, Kiran Jotheeswaran, Krishna Dhasmana
Teaching Assistant: Shravan Suravarjjala

DELTA®

ABSTRACT

- **Purpose:** Predict maintenance cases' difficulty levels using customer text reviews from Delta Faucet's survey response database.
- **Importance of Research:**
  - Allows Delta Faucet to better identify common problems with their customers' maintenance cases with their current dataset.
  - Identified several key areas Delta Faucet could improve in their survey process when attempting to acquire data for maintenance case feedback.

DATA PREPROCESSING

Remove rows with NaN values — "We really appreciate the knowledgeable staff. It did take a full 2 weeks to receive the part."

↓

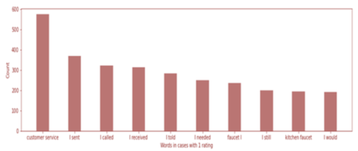
Convert Strings to Tokens — we, really, appreciate, the, knowledgeable, staff, it, did, take, a, full, 2, weeks, to, receive, the, part

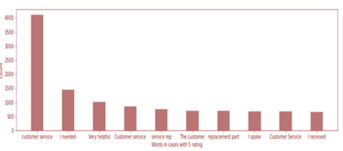
↓

Remove Stopwords — we, really, appreciate, the, knowledgeable, staff, it, did, take, a, full, 2, weeks, to, receive, the, part

↓

Populate N-grams — [really, appreciate, staff, take, full, 2, weeks, receive]





TOKENIZATION

- **Jaccard Distance Method** is effective at predicting correct spelling of words by comparing 2 Q-grams of correctly spelled word (A) with misspelled word (B).
- Two NLTK libraries (Brown/words) were compared.
- Common words were identified correctly, while uncommon words were not identified correctly.
- The prediction accuracy was greatly affected by the library of correct words.

$$d_J(A,B) = 1 - \frac{|A \cap B|}{|A \cup B|}$$

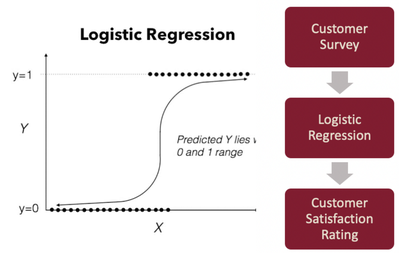
$d_J(A,B)$ **Jaccard Distance** of words A and B

$\frac{|A \cap B|}{|A \cup B|}$ **Jaccard Index (Similarity)**: number of letters appearing in both A and B divided the total counts of letters in A and B.

Tokens	Spell-corrected Tokens
niickle	nickel
profesdional	professional
fauset	faucet
4wks	40-grain
Candice	Candide
RP75675	RPM

MACHINE LEARNING

- The models 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
- Random Forest represents a **group of decision trees**.

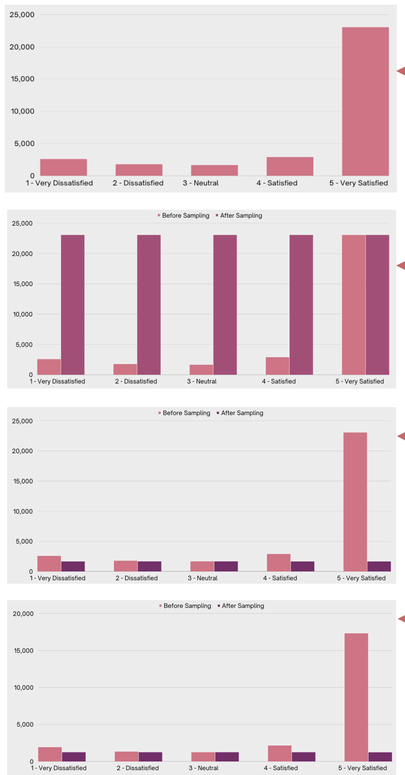


TEXT VECTORIZATION

Text Vectorization is a method of **associating textual data to a numerical score**. This step is crucial as models can only interpret numerical values for inputs. We focused on two main methods:

1. **TF-IDF**: An easy to implement vectorizer but not as powerful as Word2Vec
2. **Word2Vec**: A slightly more complicated routine to implement but far more powerful in determining an importance of a word based on its context. It accounts for the presence of stop words.

SAMPLING METHODS



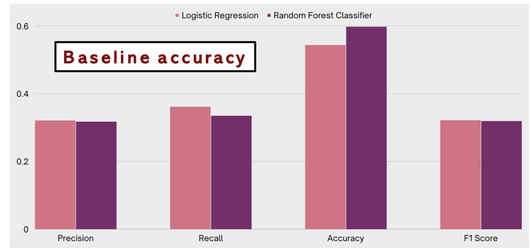
We found that we had many more samples of class "5 – Very Satisfied" than the others. This indicates that we are dealing with a dataset with class imbalance.

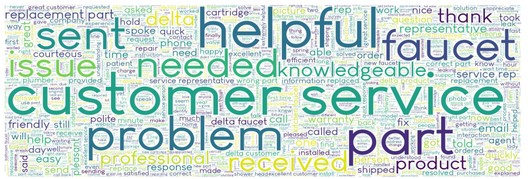
- Using the Python library **'imbleam'**, we implemented 3 potential fixes to this issue:
  - **Fix 1: Oversampling** methods copy examples in the minority class (all other than class 5 in our case) or synthesize new examples from examples in the minority class. We primarily used **Borderline SMOTE oversampling**.
  - **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**.
  - **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.

PERFORMANCE ANALYSIS / CONCLUSION

- We discovered that when using a combination of sampling techniques, the Logistic Regression model has a tendency to overfit.
- This causes **lower performances** in precision, accuracy, and recall.
- Using a GridSearchCV pipeline, we found there is **no statistically significant difference** between using different sampling techniques on our dataset.
- 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.





Dataset Word cloud

FUTURE and NEXT STEPS

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

ACKNOWLEDGEMENTS: To the **DFC** 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.