Project Report
Predicting Order Cancellations

1. Problem statement

The company's canceled orders in Q2, 2022 amounted to 9% of total sales. Order cancellations not only 'steal' revenue the company would otherwise have kept, but also disrupt the fulfillment process and result in additional costs. The benefits of being able to predict which orders will be canceled are as follows:

- It would help reduce costs associated with processing canceled orders (by placing them at the end of the fulfillment queue and/or by implementing proactive measures to prevent cancellations).
- It would allow for improved operational efficiency: fulfillment process optimization and better inventory management.
- Understanding what factors affect users' decision to cancel the order may help the business improve its product offerings, marketing strategy, and customer service.

2. Data wrangling

Data used: Amazon Sale Report for India market for Q2, 2022:
https://www.kaggle.com/datasets/thedevastator/unlock-profits-with-e-commerce-sales-data/data

Target feature: Status.
It's not clear which feature is better suited as a target feature: 'Status' or 'Courier status'. These two features have different status classifications, and there are some discrepancies between the two (an order may have 'Status': 'Shipped' and 'Courier status': 'Cancelled'). Ideally, I would talk to the data owner to understand the differences. For this project, I decided to use 'Status' as my target feature. I grouped the values into 2 main groups:

- **Shipped** - included the values: 'Shipped', 'Shipped - Damaged', 'Shipped - Delivered to Buyer', 'Shipped - Lost in Transit', 'Shipped - Out for Delivery', 'Shipped - Picked Up', 'Shipped - Rejected by Buyer', 'Shipped - Returned to Seller', 'Shipped - Returning to Seller', 'Shipping'**.**
  Note on returned orders: while the outcome of an order being returned is also not desirable for business, I assume that returns and cancellations are two different business problems and the reasons for rejecting/returning the order are different from reasons for canceling the order. Hence, keeping these orders under the 'success' ('Shipped') label for this project.
- **Cancelled.**
- I dropped the records with 'Pending' & 'Pending - Waiting for Pick Up' status since these orders' final status is unclear.

Data cleaning steps:

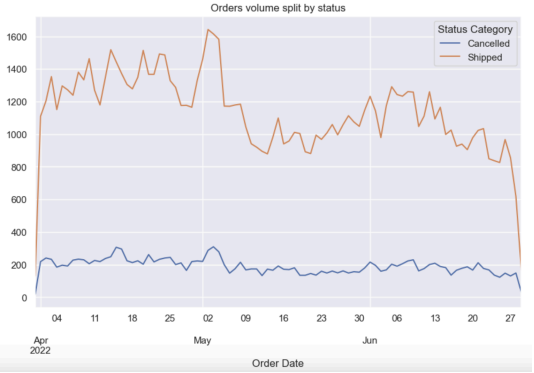
1. I dropped the 'Qty' column, since for canceled orders the quantity was getting recorded as 0, so using the column is not informative for predicting cancellations.
2. Values for the amount were also missing for canceled orders. Since I want to use product price in my model, I imputed price based on the closest values for the same SKUs that are not null. I didn't want to impute with average value, because at different periods same product could have different prices (possibly, due to promotions).
3. I standardized values for states. The initial dataset had 69 unique values (while India has 36 states & territories). I lowercased the values, and used the list of India's states from Wikipedia to fuzzy match it with states from the dataset. Additionally, I made a few manual changes to the values that didn't get matched.
4. I dropped one duplicated Order id.
5. I split the 'Promotions' column into separate columns for different types of promotions: 'Free financing', 'Free shipping', 'Coupon', 'Other promotions', and 'No promotions'.

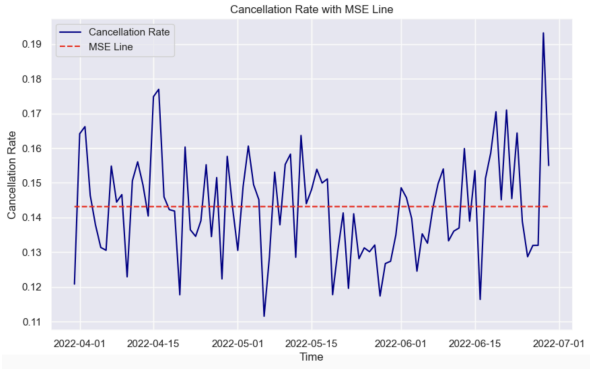
3. Exploratory Data Analysis

3.1. Exploring target feature

My target feature has unbalanced classes - only 14% of observations have status = 'Canceled'.

I checked orders volume split by status and cancellation rate trend - while orders volume has dropped in May, there doesn't appear to be a change in cancellation rate trend.



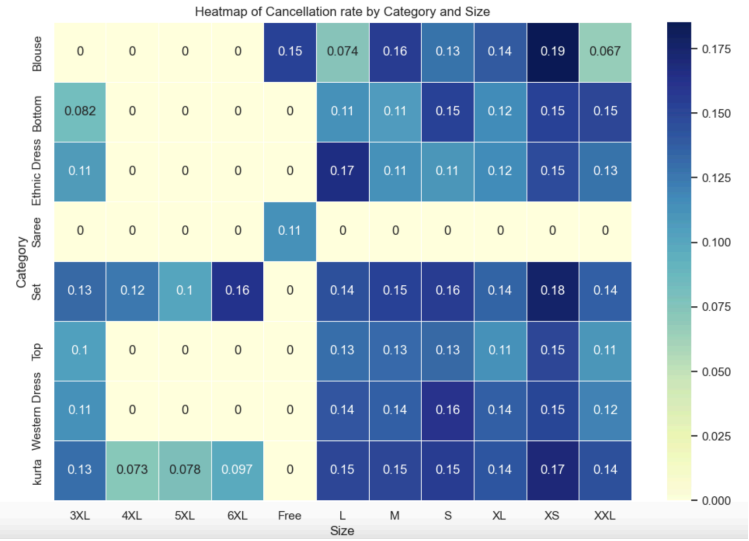


3.2. Exploring categorical features

I wrote a function to calculate and visualise cancelation rates based on different dimensions, and to calculate chi-square statistics to make a conclusion of how likely the difference in cancellation rate between classes is due to chance.

Findings:

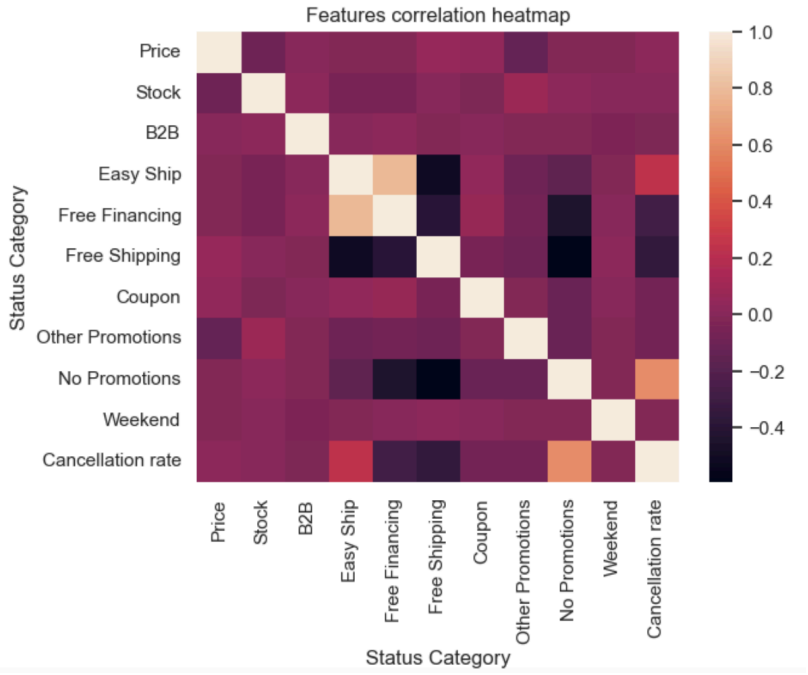
- Some product categories, SKUs, and sizes have significantly higher cancelation rate.



- Some shipment states have significantly higher cancelation rates.

3.3. Exploring numerical features

I created a correlation heatmap to explore the relationships of numerical features between themselves and with the target feature.



'Easy ship', 'Free financing', and 'Free shipping' seem to have the strongest negative correlation with cancelation rate, while 'No promotions' has a positive correlation.

Looking at scatterplot of cancelation rate vs price split by promotions - there's a much lower cancelation rate (often times it's 0) for orders with promotion. This could indicate either customers' higher satisfaction with purchase when there's a promotion, or terms of promotion itself (returns not allowed).

I also checked 'Price' feature for outliers - there are some, so I'll need to deal with them during modeling step.

4. Pre-processing

4.1. I created new features:

- 'Month'
- 'DayofMonth'
- 'begin_of_month'
- 'midddle_of_month'
- 'end_of_month'

4.2. I transformed the target feature to numerical ('Cancelled' = 1).

4.3. I one-hot encoded categorical features: 'Category', 'Size', 'Shipment State Matched', 'Shipment Type'.
4.4. Due to the huge number of unique SKU values I didn't want to one-hot encode this feature not to explode the dimensionality of my data. I target encoded 'SKU' feature instead.
4.5. I scaled numerical features 'Price', 'Stock', 'DayofMonth' and 'Month'. I used RobustScaler to deal with outliers.
4.6. I split the data into train and test.

My final list of features is below:

- Price (numerical)
- Stock (numerical)
- DayofMonth (numerical)
- Month (numerical
- SKU (target encoded)
- B2B (bool)
- Easy Ship (bool)
- Free Financing (bool)
- Free Shipping (bool)
- Coupon (bool)
- Other promotions (bool)
- No promotions (bool)
- Category (one-hot encoded)
- Size (one-hot encoded)
- Shipment State (one-hot encoded)
- Shipment Type (one-hot encoded)

5. Modelling

Since the target feature has imbalanced classes, 'accuracy' is not a good metric to assess model performance. I used 'recall' to evaluate how well the model predicts cancellations.

I tried 4 classification models:

- Logistic regression.
  I used *class_weight = 'balanced'* parameter to accommodate for imbalanced classes of target feature.

- KNN Classifier.
  I resampled my data set using *SMOTE()* to accommodate for imbalanced classes of target feature. I tried *n_neighbours* = 3, 5, 7, 9 & 11 and used *n_neighbours = 11* in my final KNN model since it had the highest recall.

- Random Forest
  I used *class_weight = 'balanced'* parameter to accommodate for imbalanced classes of target feature.

- Decision Tree Classifier
  I used *class_weight = 'balanced'* parameter to accommodate for imbalanced classes of target feature. Also, I used unscaled features to fit the model, since decision trees don't rely on distance for predictions, and unscaled features are easier to interpret.

Classification report for the models

Model	accuracy	precision (for label 1)	recall (for label 1)	f1-score (for label 1)	support (for label 1)
Logistic regression	0.81	0.42	0.94	0.58	5077
KNN Classifier	0.80	0.40	0.84	0.55	5077
Decision Tree	0.76	0.36	0.98	0.53	5077
Random Forest	0.91	0.77	0.55	0.64	5077

Decision tree produced the best recall - 0.98 on the test data. Hence, I chose Decision Tree as my final model for predictions. I used Bayesian optimization to tune the following hyper-parameters for my model:

- max_depth,
- min_samples_split,
- min_samples_leaf,
- criterion,
- max_features,
- max_leaf_nodes.

However, the optimized parameters (max_depth=6, criterion="gini", max_leaf_nodes=58, min_samples_leaf=47, min_samples_split=71, max_features = 'sqrt') didn't change my model performance.

After fitting the Decision tree model with optimized hyperparameters and stratified cross-validation the final Classification report for the created model is below.

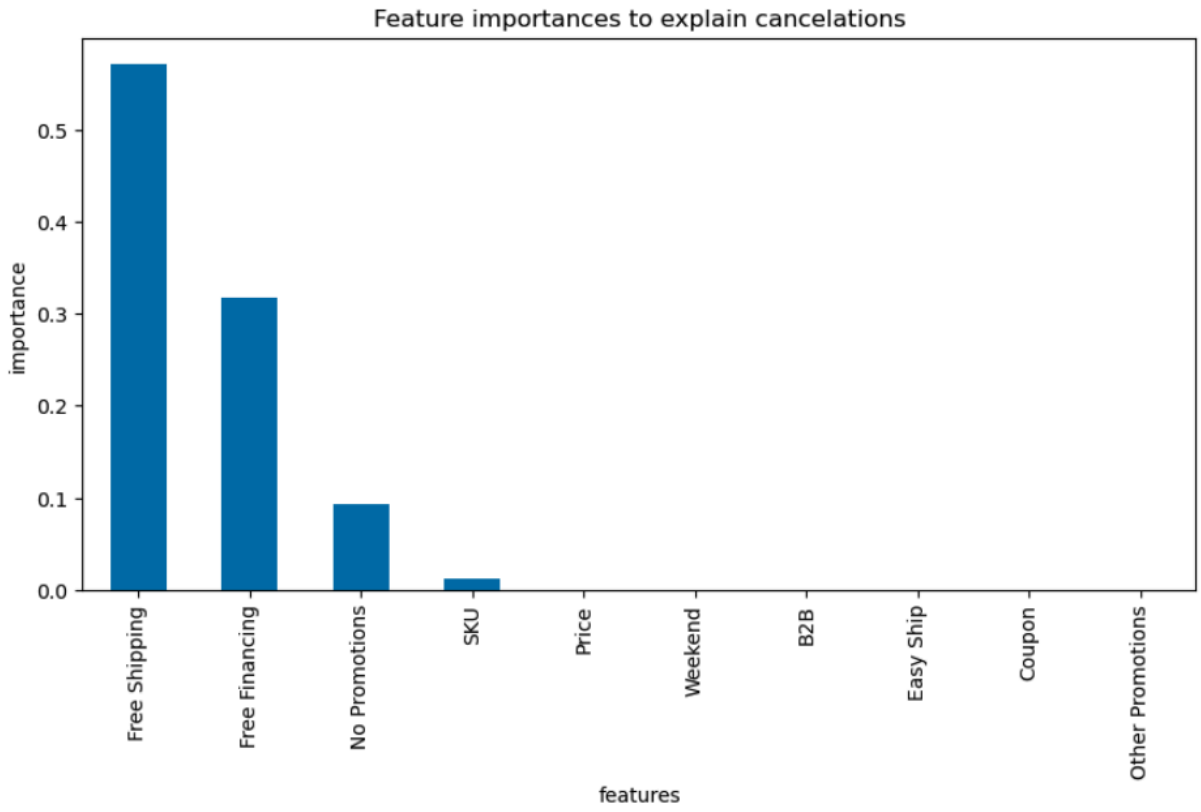
```
Classification report:
              precision    recall  f1-score   support

           0       1.00      0.64      0.78     31007
           1       0.31      0.98      0.47      5077

    accuracy                           0.69     36084
   macro avg       0.65      0.81      0.62     36084
weighted avg       0.90      0.69      0.73     36084
```

6. Takeaways

Looking at feature importances, based on which my model is predicting cancellations, the most important ones are: Free shipping, Free financing, No promotions & SKU.



Orders with Free shipping are associated with fewer cancellations. An intuitive explanation could be that customer satisfaction is higher for orders with free shipping due to the higher perceived value of the order.

Orders with Free financing are associated with fewer cancellations. A possible explanation could be also tied to customer satisfaction, or maybe free financing has some rules (like, returns are not allowed or the terms are more strict).

Orders with no promotion have a higher cancellation rate. As opposed to orders with promotions, orders without promotions may have lower perceived value and result in lower customer satisfaction.

Some SKUs have higher cancellations. This could be due to concerns about the quality of some products (perhaps, the website has some bad reviews which make customers hesitate and cancel their order), or maybe some SKUs are out of stock or the fulfillment process takes more time.

7. Future research

It would be good to incorporate some fulfillment aspects of each order into the model - like fulfillment time (for the orders with the status ‘Shipped’) and the time elapsed between the order

and cancellation (for orders with the status ‘Canceled’) since long wait times could be one of the reasons for cancellations.

It would be good to look at some customer characteristics - age, gender, first-time customer vs repeat customer, time spent on the website before placing an order - since customer characteristics and spontaneity of purchase could also have an impact on the propensity to cancel the order.