The goal is to Identify main reason for customer satisfaction or dissatisfaction

Predict satisfaction level by category and product type

The client is Olits, Brazilian version of Amazon.com, they see satisfaction scores for purchases but not sure what are exact reasons, the order review text is pretty short and there are around 9000 of them. Olits would learn what are the reasons for satisfaction or dissatisfaction and make suggestions to their sellers accordingly.

The data is posted on Kaggle

I decided to take a route of supervised learning , after labeling data for different reasons of satisfaction or dissatisfaction, I want to classify it and predict label for the rest of the data

The goal is to create a model that predicts a reason for a bad or good score, a model that predicts what categories give a good or bad score

*The goal of this project is to find out what are the reasons*

*customers are satisfied or dissatisfied with their purchase.*

*Because of the nature of the data given to answer that question, 2 algorithms has be in place:*

*Identifying a topic of each given review*

*Finding features that make a review positive, negative , and fall under a certain topic*

The first challenge is to identify review topic.Statement analysis showed that topics of the review could be - delivery, quantity, quality, wrong item delivered, customer service etc. After common topics were identified reviews were manually labeled.

Preprocessing:

This part of the project requires only one of many table provided by olits: reviews. Review consists of the the review title, review, and the score, review and it’s title concatenated (it is a required step as sometimes customer put the reason for the review in the title: Title: Fast Delivery Review: I am so happy, Thank you), Words are stemmed ,Most informative features are extracted. Examining those words can help identify main reasons for satisfaction or dissatisfaction for the order

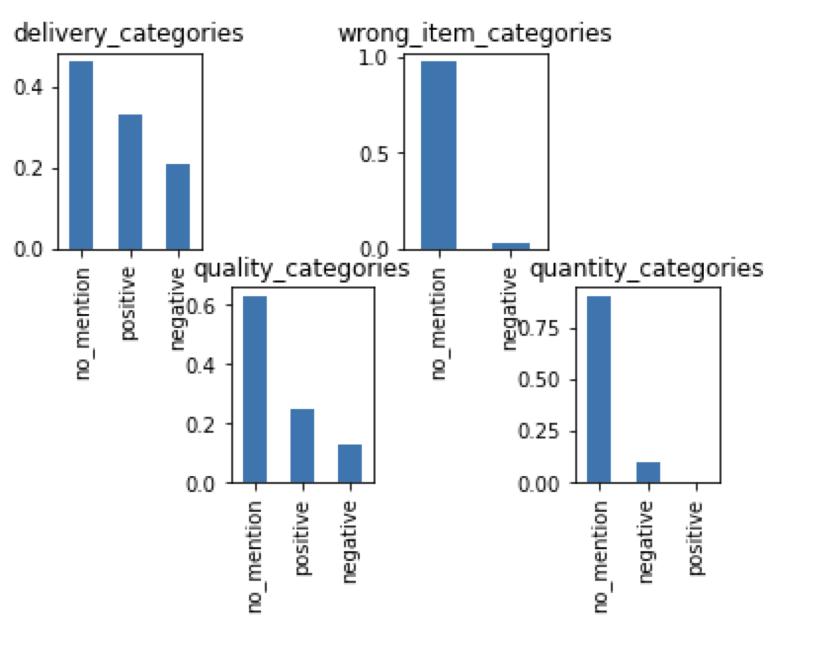
Labeling:

Using exploratory sentiment analysis revealed few main reason for satisfaction/dissatisfaction:

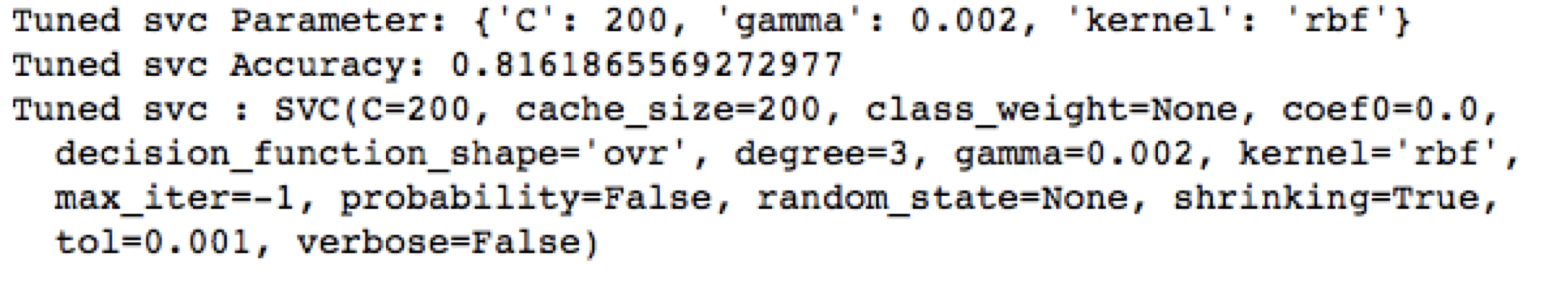
* - Quality of a product(s) - sentences describing color, size, dimensions, aesthetic qualities indicate that review is about quality, it could be both positive and negative.
* - Quantity of product(s) - lack of full order. Mostly in negative tone, this reviews are usually complaints about the difference between the number of products that was ordered and the number of products that was delivered.
* - Delivery - There are a lot of verbs describing delivery , verb 'delivered' (entregue) and others may not signify any connotation and may appear in other contexts. And there adverbs that refer to timing that actually means that the comment is about the delivery.
* - Customer service - there are some verbs that may refer to customer service: email, phone.

While this sentiment analysis may show how strong a word is associated with the reason for satisfaction or dissatisfaction with an order.

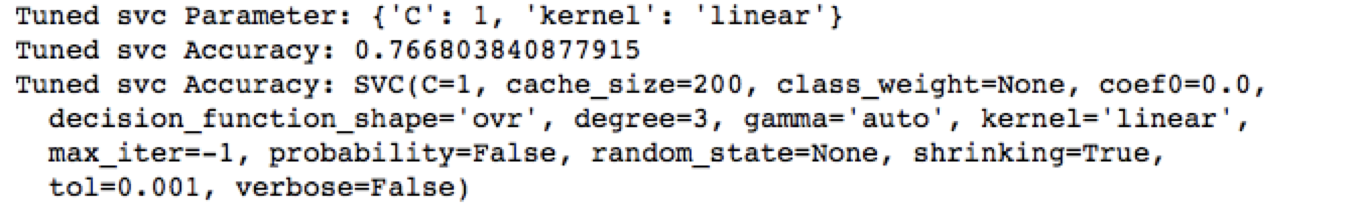
Labeling of randomly selected data gave following value counts :



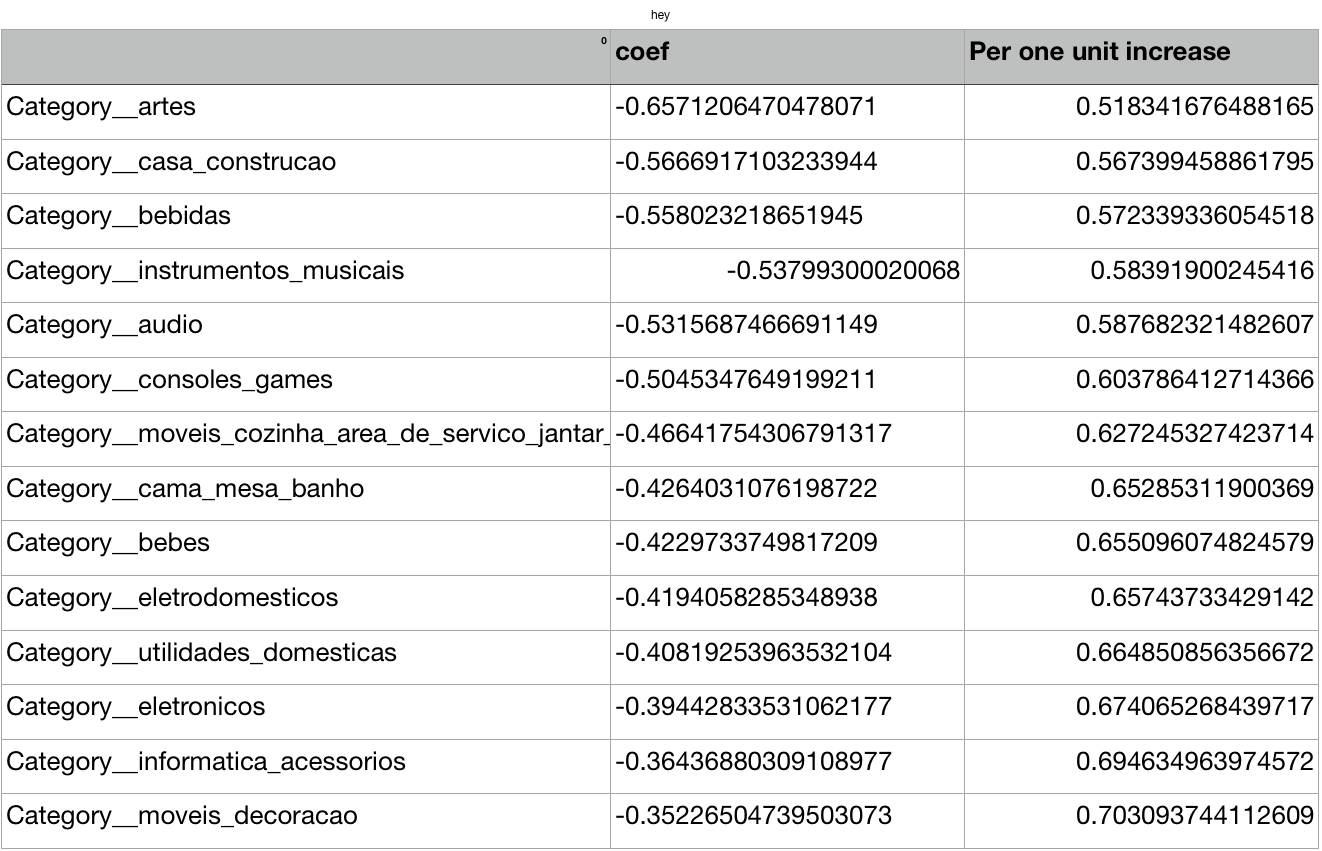
After Data is labeled data was put through algorithm, to see what is the topic of the reviews. Using implementing gridsearch and model selection algorithm following is the is the best model to predict delivery topic:



Using implementing gridsearch and model selection algorithm following is the is the best model to predict quality topic:



Olits provides a lot of tables that need to be joined, to use them for the next model:

* Categories were ‘pivoted’ (one hot encoding)
* Engineered features : delivery in hours (calculated from deducting delivery time from order time)
* Tables were joined
* New columns of predicted review about delivery and predicted review about quality were added
* 

This fitted model says that, holding every variable at a fixed value, the odds of getting positive review on delivery in category artes over the odds of not getting a negative one is exp(.65) = 0.51. In terms of percent change, we can say that the odds for getting god review is 51% higher than getting bad review.

Best estimator for quality reviews about quality was :

est Estimator

RandomForestClassifier(bootstrap=True, class\_weight=None, criterion='entropy',

max\_depth=None, max\_features=6, max\_leaf\_nodes=None,

min\_impurity\_decrease=0.0, min\_impurity\_split=None,

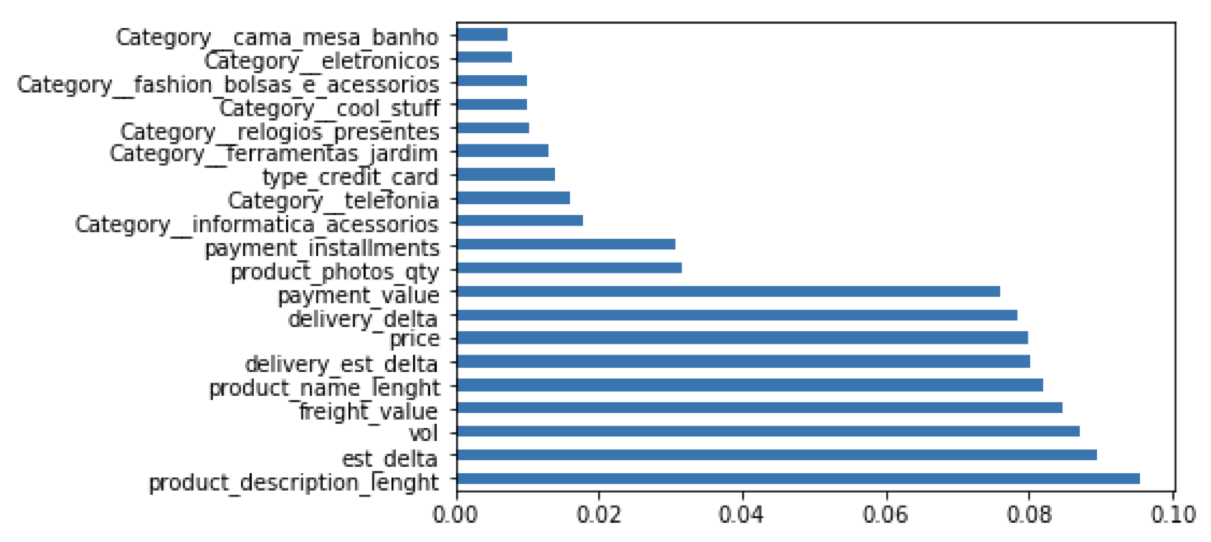
min\_samples\_leaf=6, min\_samples\_split=2,

min\_weight\_fraction\_leaf=0.0, n\_estimators=10, n\_jobs=1,

oob\_score=False, random\_state=None, verbose=0,

warm\_start=False)

Best features:



Recommendations:

* Although review comments on delivery is the most common topic. There are very weak coefficients across the board.
* Not surprisingly Product description and number of pictures play a role in a good review. Olits could recommend sellers to give more information about the product, so that the customer knows what to expect and more likely to give a good review.
* Estimation of product arrival and the arrival date of the product actually matters for perception of quality. It looks like delivering on time makes customers give a positive review about the quality.
* Price have positive impact on quality perception. Customer wouldn’t mind to praise a good, expensive product (as long as it is good quality)

Limitations:

* Originally manual labeling found more reasons for good or bad reviews :
  + For example : wrong product, good customer service, or costumer service is not answering at all, product was not delivered at all, or quantity was wrong, although it is not the most common problems, labeling more data could help with identifying the main reason for such problems.
* Reviews were translated by google translate, some topics may have been missed or mislabeled
* Project consists of 2 algorithms, which decreases overall accuracy