\* In order to save time, we sampled 12,000 of our original 114,066 records.

\* We utilized and 80/20 test-train split

\* We used intuitive feature reduction prior to training the model to reduce computational resources and time and put some thought into what features would ideally contribute to a product arriving by it’s estimated arrival date. While it may seem intuitive to just assume comparing the estimated arrival date and the actual arrival date would suffice, we thought about how people perceive estimated delivery times from couriers. For example, if I can pay for a package to be overnighted from Dallas to Austin and pay a premium rate; however, I also know that generally I can just pay the ground shipping rate the and the package will show up with 1-2 days rather than the estimated 5-7 days, so that is why we wanted to really have features that contributed to determining if a package was truly late, on time or early.

\* We ran 2 basic models using only minimal tuning for multinomial classification and 2 refitted models using RandomizedSearchCV for tuning.

\* We used multinomial classification to classify if packages were late, on time or early to olist customers

\* Basic Logistical Regression is the only model that was able to categorize deliveries that were made on-time (value #2) with a precision of 6%. It also had precision of 26% for late deliveries and 89% for early deliveries which makes sense since early deliveries had the largest representation in the dataset, followed by late deliveries and then on time deliveries. The overall precision was 75%.

\* Refitted using RandomizedSearchCV Logistical Regression: Once we tuned the model using RandomizedSearchCV with a 5-fold validation, the best model also correctly classifies on time deliveries 6% of the time, late deliveries 32% of the time and early deliveries 89% with and overall accuracy of 87%. So we saw some improvement with correctly categorizing the late deliveries and the overall accuracy.

\* SVM with Stochastic Descent with minor tuning to handle multinomial classification returned an overall accuracy of 87%, but failed to correctly classify any on time packages. This model had the same precision for late deliveries – 26% and early deliveries- 89%, but saw an increase in the recall for early deliveries, signaling an overfit at 98% with only 9% recall for late deliveries and no recall for on time deliveries. The overall accuracy was 87% and on par with the overall accuracy from the tuned logistal regression model (also at 87%).

\* Reffited SVM with Stochastic Descent: we again utilized RandomizedSearchCV to tune our SVM with stochastic decent that maintaining an overall accuracy of 86% that was slightly lower than the untraining SVM (87% accuracy) and the training logistical regression model (87%). The precision was better for categorizing late deliveries (30%) compared to the untrained SVM (26%), but lower than the training logistic regression model (32%); however, the recall for late deliveries classification was the highest for the trained SVM (17%) compared to the 2nd best recall for late deliveries (15%) from the untrained logistical regression model.

\* Time - we found the best time for a model was the RandomizedSearchCV trained SVM at 0.06 seconds compared to 0.33 seconds for the best trained logistical regression model that also used RandomizedSearchCV.

\* To get a sense of how the untrained models performed we looked at the wall times using the %%time magic function and the fitted svm was best at 0ns while the worst was the fitted logistic regression at 7.09s. The untrained logistic regression ran at 349ms and the untrained SVM ran at 51.5ms

\* The overall best model in terms of time, overall accuracy was the trained SVM model, but due to its lack of ability to correctly classify on time deliveries we are leaning towards the untrained logistical regression model to ensure we have some representation from the model.