Machine Learning Engineer Nanodegree

Capstone Project

Adarsha Badarinath

April 2017

NOTE: A capstone proposal has already been accepted. Here are the links to proposal and the review of the proposal. A copy of the Kaggle data can be found here for the purpose of verification in-case the original data changes.

I. Definition

Project Overview

Based on historical data predict backorder risk for products. Part backorders is a common supply chain problem. Working to identify parts at risk of backorder before the event occurs so the business has time to react.

What is backorder?

Popular items may sell out quickly and temporarily be on backorder. This means that the items are currently out of stock but that there are shipments on their way to re-stock our warehouses.

Why is it helpful to predict it??

Whenever a backorder situation occurs it is very easy for a competitor to woo the customer to their product and even sell it at a premium taking advantage of the scarcity. Also sometimes businesses have to give discounts to retain the customer. It affects the bottomline of any business.

Dataset I use is part of the recently released kagel dataset.

Problem Statement

> *Based on historical data predict backorder risk for products. The data is taken from kagel dataset. Using the dataset we predict 'went_on_backorder' column value given other input parameters. Explore different methods and different*

techniques to solve the problems. We want to answer the following questions on the dataset in the end

1. Given all the columns in the dataset predict ‘went_on_backorder’ column. Which technique is the best to do this?
2. Find out which are the top 5 features which has the most impact on our predictions

Inputs:

Input data format is in the form of a CSV

- sku - Random ID for the product - Integer
- national_inv - Current inventory level for the part - Integer
- lead_time - Transit time for product (if available) - Integer
- in_transit_qty - Amount of product in transit from source - Integer
- forecast_3_month - Forecast sales for the next 3 months - Integer
- forecast_6_month - Forecast sales for the next 6 months - Integer
- forecast_9_month - Forecast sales for the next 9 months - Integer
- sales_1_month - Sales quantity for the prior 1 month time period - Integer
- sales_3_month - Sales quantity for the prior 3 month time period - Integer
- sales_6_month - Sales quantity for the prior 6 month time period - Integer
- sales_9_month - Sales quantity for the prior 9 month time period - Integer
- min_bank - Minimum recommend amount to stock - Integer
- potential_issue - Source issue for part identified - Boolean
- pieces_past_due - Parts overdue from source - Integer
- perf_6_month_avg - Source performance for prior 6 month period - Integer
- perf_12_month_avg - Source performance for prior 12 month period - Integer
- local_bo_qty - Amount of stock orders overdue - Integer
- deck_risk - Part risk flag - Boolean
- oe_constraint - Part risk flag - Boolean
- ppap_risk - Part risk flag - Boolean
- stop_auto_buy - Part risk flag - Boolean
- rev_stop - Part risk flag - Boolean

Output:

Yes - Product went on back order

No - Product did not go to backorder

Machine learning task:

- The main task is binary classification problem between Yes/No of predicting if an item went on backorder
- To determine if this is a good enough indicator to take business decisions on.

Metrics

I am using the evaluation metric of Precision as my primary evaluation metric. Along with this I will consider the F-1 score.

$$\text{Precision} = \frac{tp}{tp + fp}$$

Tp - True positive. | fp - False positive

$$\text{Recall} = \frac{tp}{tp + fn}$$

Fn - false negative

$$F_1 = 2 \cdot \frac{1}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Why?

In my usecase predicting a backorder places a realworld order to be dispatched to the wearhouse/shop. This means we will produce a product and have revenue impact. We should not do this unless absolutely required. Hence precision determines this. We select f1-score as secondary as it captures the overall picture including the recall score.

II. Analysis

Data Exploration

Our input has a highly structured data in a CSV/Excel format. The data also is quite clean and doesnt have any null/bad values.

The columns or features of this data set are

['national_inv' 'lead_time' 'in_transit_qty' 'forecast_3_month'

'forecast_6_month' 'forecast_9_month' 'sales_1_month' 'sales_3_month'

'sales_6_month' 'sales_9_month' 'min_bank' 'potential_issue'

'pieces_past_due' 'perf_6_month_avg' 'perf_12_month_avg' 'local_bo_qty'

'deck_risk' 'oe_constraint' 'ppap_risk' 'stop_auto_buy' 'rev_stop'

'went_on_backorder']

The total number of rows is = 1693050

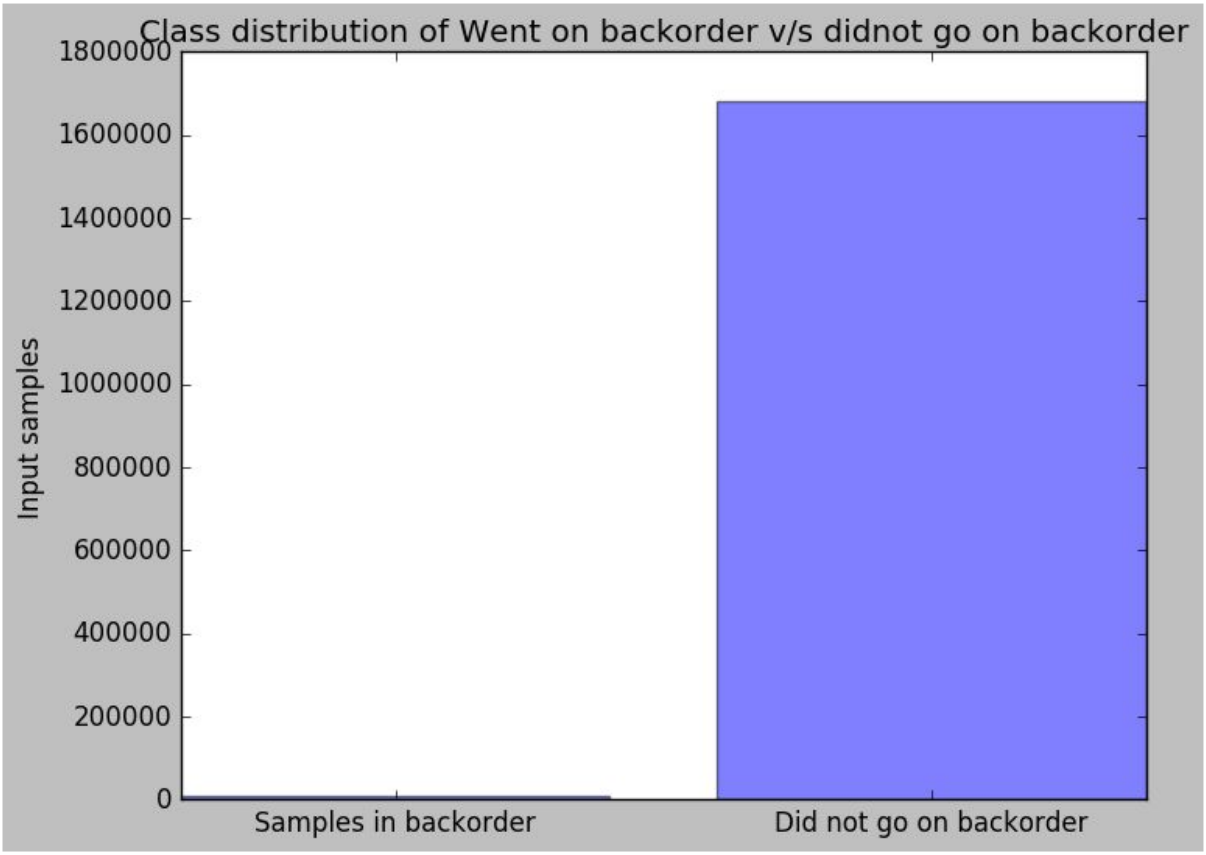
We have 7 columns (including went_on_backorder) which are binary columns with Yes/No value. The rest of the 14 are integer/floating point numbers.

The SKU is a unique ID in the data and we can use it as a row identifier.

Exploratory Visualization

Lets examine the class distribution of 'went_on_backorder' class we are trying to predict.
went_on_backorder = Yes = (10914, 22)
went_on_backorder = No = (1682136, 22)
went_on_backorder = % of Yes/no = 0.6446354212811198

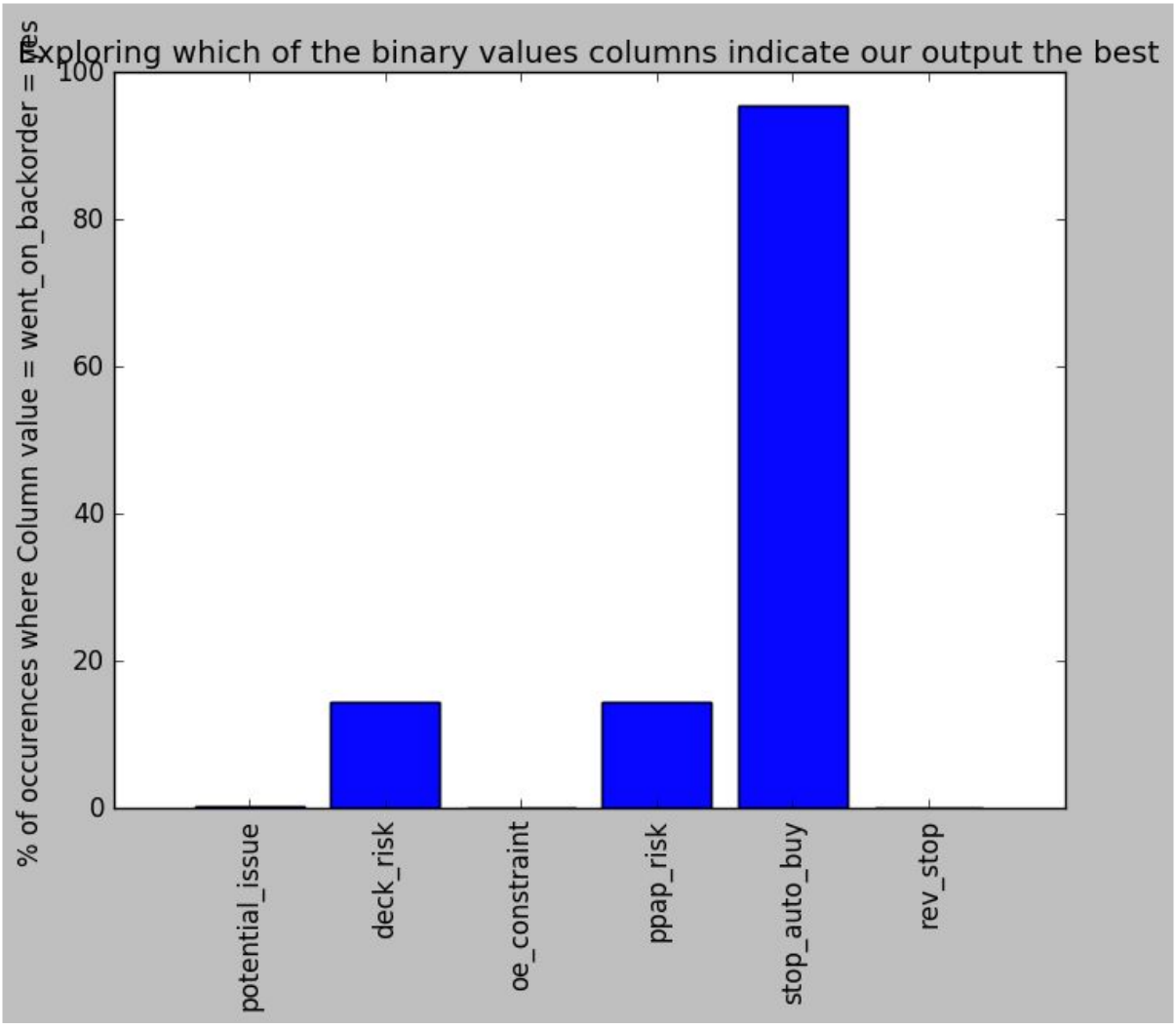
We see a distribution of <1% on the entire data for Yes v/s No. Hence its a very highly skewed distribution.



The above provides insight that the data we are exploring is highly skewed. This helps us tune our algorithms later like Neural networks where we have to account for this imbalance and drop majority of data who's went_on_backorder = false from the data. This helps us not to overfit the data for a single class.

Let us also examine how much of a role the binary columns play. Below are the binary columns whos' value can take YES/NO.
['potential_issue','deck_risk','oe_constraint','ppap_risk','stop_auto_buy','rev_stop']

If we blindly predict went_on_backorder = yes when the a binary column = YES, we see that 95.6% of the time stop_auto_buy is true. However this is one of the known part risk flags. Logically thinking whenever we stop auto ordering more inventory, naturally we are susceptible to backorders. Other interesting feature is deck_risk and ppap_risk are around 14%.



Algorithms and Techniques

Decision Trees

This modal helps in choosing the feature with the highest information and making a branch. It does it again with the remaining features until we run-out-of features or a specific depth. A decision tree has only burst nodes (splitting paths) but no sink nodes (converging paths).

There are several parameters/tuning options.

- Which “feature” do we expose to the tree in the data
- How is “information gain” calculated
- What’s the max depth allowed

sklearn.model_selection.GridSearchCV - provides a lot more parameters and fine tuning options.

Decision trees have a tendency to overfit to the data. And *hence we will be using* ***Random Forest***. Which is an ensemble of decision trees.

Support Vector Machines

Support vector machines are machine learning algorithms which draw boundaries in N dimensional data and try to maximize this boundary gap to give the best separation.

They can be used in both supervised and unsupervised learning scenarios. For classification technique like ours it is a very good algorithm to try.

Neural Networks

They are modeled based on neurons of brain, which take multiple inputs and combines them with a probabilistic dynamic weight and sends a single output. They primarily produce a regression output but can also be used for classification purposes. The real advantage is in when we “back propagate” during training. We let the network predict based on random values at first and then go back and tune the weights based on how far off was it from the actual value. Networks can be constructed in many forms by combining many “neurons” both in depth and width.

Benchmark

NOTE: You can find all the model and data exploration in this notebook.

First model I considered:

A simple benchmark model predicting Backorder required = Yes(i.e we need to backorder) IF ‘stop_auto_buy’ column is true results in. Upon initial analysis I found this Part risk flag seems to be a good indicator.

Total size = 1693050

True positive = 10434

True negative = 65233

False positive = 1616903

False negative = 480

Precision = 0.64%

Recall = 95.6%

Accuracy = 4.4%

The above model suffers from a high recall but low precision.

Second model I considered is a “RandomForestClassifier”.

	precision	recall	f1-score	support
0	1.00	0.89	0.94	225210
1	0.07	0.82	0.13	2280
avg / total	0.99	0.89	0.93	227490

This has lower recall but higher precision and f-1 score.

III. Methodology

Data Preprocessing

1. Data is in form CSV needs to be imported.
2. Data was originally cleaned to not contain any null/empty values. No such row was found as the kaggle data was pretty well cleaned up already.
3. Certain boolean columns need to be manually converted to a 0/1 integer value from ‘Yes’/’No’ String
4. Train and test split was done and we have different CSV files written into disk.
5. We split into X , y by dropping ‘went_on_backorder’ column for X and storing this column in ‘y’ a.k.a our desired prediction

Preprocessing for Decision Trees model

- Did no preprocessing of data for and supplied the full X,y for fit() function

- The MAX_DEPTH parameter feature was tuned with the help of RandomForestClassifier estimator using feature_importances_ array.

Preprocessing for Support Vector Machines model

Apart from the standard preprocessing explained above, no special preprocessing is needed for this algorithm. We however do grid search to find best parameters which is explained later.

Preprocessing for Neural Networks model

Neural networks need to have their input regularized to avoid large variations. Thus we preprocess the data in the cells to be >= 0 and <= 1.

Neural Networks need more balanced data. Hence I dropped 80% of class=0 .. i.e where backorder = NO. This enabled a more balanced dataset where the classes are more evenly distributed. Below is the code which does it

```
for i, row in X_regularized.iterrows():
    if y_copy[i] == 0 and randint(0,10000) < 8000:
        rows_to_drop.append(i)

X_regularized.drop(rows_to_drop,inplace=True)
y_copy.drop(rows_to_drop,inplace=True)
```

Implementation and Refinement

Decision Trees

I have implemented both DecisionTreeClassifier & RandomForestClassifier (which is an ensemble of decision trees). Both required normal preprocessing explained before. Analysis was made to identify top features as shown before.

To find the best model I applied GridSearchCV to find the best parameters. In both cases StratifiedKFold split was used to make sure we are not tuning to the training data.

Tuned parameters for DecisionTreeClassifier

{'max_depth':[10,15]}

Tuned parameters for RandomForestClassifier

[{'n_estimators': [100,200],

'criterion': ['entropy','gini'],

'max_depth':[10,15]}]

For RandomForestClassifier I also hyper-tuned over score ['precision', 'f1'].

```
cv = StratifiedKFold(n_splits=5)
clf = GridSearchCV(DecisionTreeClassifier(random_state=42),
tuned_parameters, cv=cv,verbose=10, n_jobs=4)
clf.fit(X, y)
```

```
cv = StratifiedKFold(n_splits=5)clf =
GridSearchCV(RandomForestClassifier(class_weight='balanced'),
tuned_parameters, cv=cv,scoring='%s_macro' % score, verbose=10,
n_jobs=4)    clf.fit(X, y)
```

Support Vector Machines

I tried first a normal SVC() classifier from sklearn. But then I found that it is not well suited for my dataset. From sklearn documentation it says “The fit time complexity is more than quadratic with the number of samples which makes it hard to scale to dataset with more than a couple of 10000 samples.”.

Thus I used a LinearSVC which scales well with more data. My input were for the GridSearchCV was as follows

tuned_parameters = [{'C' : [0.5,1.0,1.5],

'loss' : ['hinge','squared_hinge']}]

scores = ['precision', 'f1']

I choose the above tuned_parameters on a few experimentation. I then choose to choose the “scoring” function as ‘precision’ and ‘f1’ as those are my most important parameter to tune to.

```
cv = StratifiedKFold(n_splits=4)
clf = GridSearchCV(LinearSVC(random_state=42), tuned_parameters,
```

```
scoring='%s_macro' % score, cv=cv, verbose=10, n_jobs=8)
clf.fit(X, y)
```

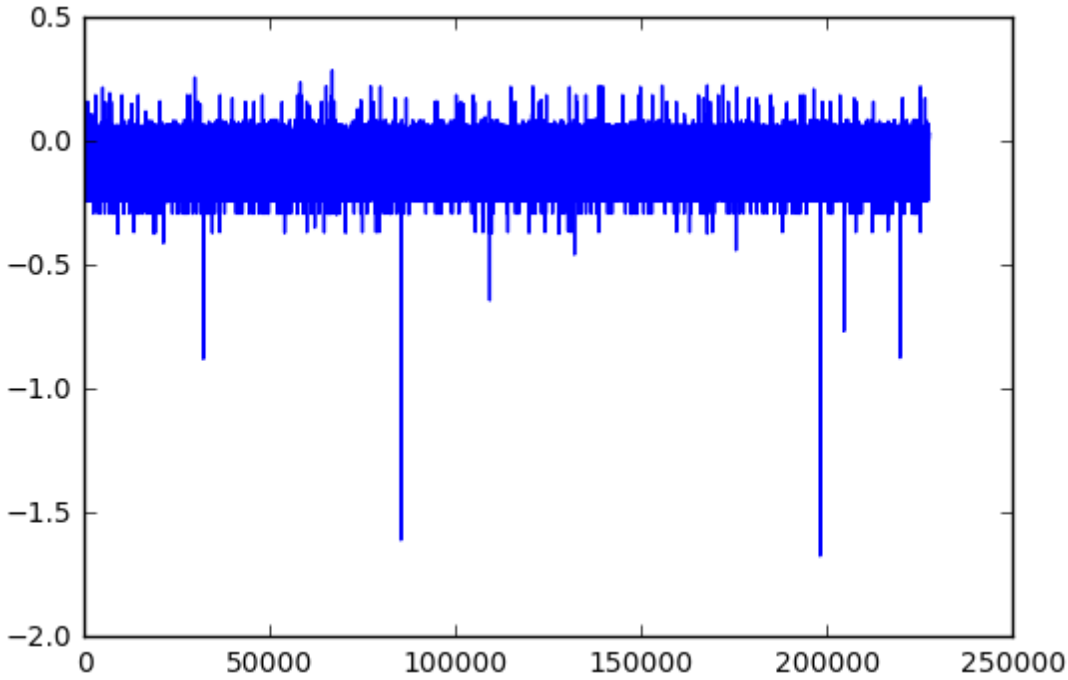
Neural Networks

I used keras front end for Neural Networks to run on a TensorFlow backend on AWS graphics card for fast iteration and experimentation.

Below is my model summary

```
Layer (type)                 Output Shape              Param #     Connected to
====================================================================================
dense_51 (Dense)             (None, 21)                462         dense_input_17[0][0]
____________________________________________________________________________________
activation_34 (Activation)   (None, 21)                0           dense_51[0][0]
____________________________________________________________________________________
dense_52 (Dense)             (None, 14)                308         activation_34[0][0]
____________________________________________________________________________________
activation_35 (Activation)   (None, 14)                0           dense_52[0][0]
____________________________________________________________________________________
dense_53 (Dense)             (None, 7)                 105         activation_35[0][0]
____________________________________________________________________________________
dense_54 (Dense)             (None, 1)                 8           dense_53[0][0]
====================================================================================
Total params: 883
Trainable params: 883
Non-trainable params: 0
____________________________________________________________________________________
```

I trained it for 4 Epoch and batch_size=6400. Finally I predicted on the Regularized TEST data.
I plotted the predicted regularized output for classes. Since NN predict a value between 0 & 1, I had to draw this choose a value below/above which we predict backorder=true.



From above I arrived at

```
if (classes[i]) > 0.1 or classes[i] < -0.35:
    classes_pred[i] = 1
else:
    classes_pred[i] = 0
```

IV. Results

Model Evaluation and Validation and Justification

(Green - better than benchmark. Darker the better)

	Best params (if available)	**Precision**(0/1 & Total)		**F1 score** (0/1 & Total)	
Benchmark model					
		1.00 / 0.07	.99	0.94 / 0.13	.93
Decision Trees					
Decision Tree	{'max_depth': 10}	0.99 / 0.29	.98	0.99 / 0.03	.99
Random Forest	{'criterion':	1.00 / 0.10	.99	0.97 / 0.18	.96

(Precision tuned)	'entropy', 'max_depth': 15, 'n_estimators': 100}				
Random Forest (f1 tuned)	{'criterion': 'entropy', 'max_depth': 15, 'n_estimators': 200}	1.00 / 0.10	.99	0.97 / 0.18	.96
Support Vector Machines					
Precision tuned	{'loss': 'hinge', 'C': 1.0}	0.99 / 0.02	.98	0.99 / 0.00	.98
F1 tuned	{'loss': 'squared_hinge', 'C': 1.0}	0.99 / 0.05	.98	0.99 / 0.06	.98
Neural Network					
		0.99 / 0.05	.98	0.99 / 0.01	.98

I have explained why Precision is our main criteria followed by f1 score in the previous sections

The above numbers were generated through stratifiedKfold techniques where possible. This will make sure we do not optimize on the training data and get correct robust result results.

Looking at the results above, Random Forest are the clear winner. A simple decision tree with Max-depth = 10 is also better than the benchmark model. Random forest are better at avoiding overfitting than simple decision tree. Thus we see overall F1 and precision is better for random forest

NOTE: even tho for precision Random forest is 4x better at prediction of a product going into backorder, we see the f1 score is significantly down (3x down) due to overfitting

The final solution is **42% better precision and 38% better f1 score** than benchmark in predicting the product went on backorder (class = 1). Also for predicting NOT-going-to-be-backordered there is no degradation in precision and 3% increase in f1 score.

Overall I would recommend the any company on benchmark models switch over to this modal to decrease the chance of products being not available for being on backorder.

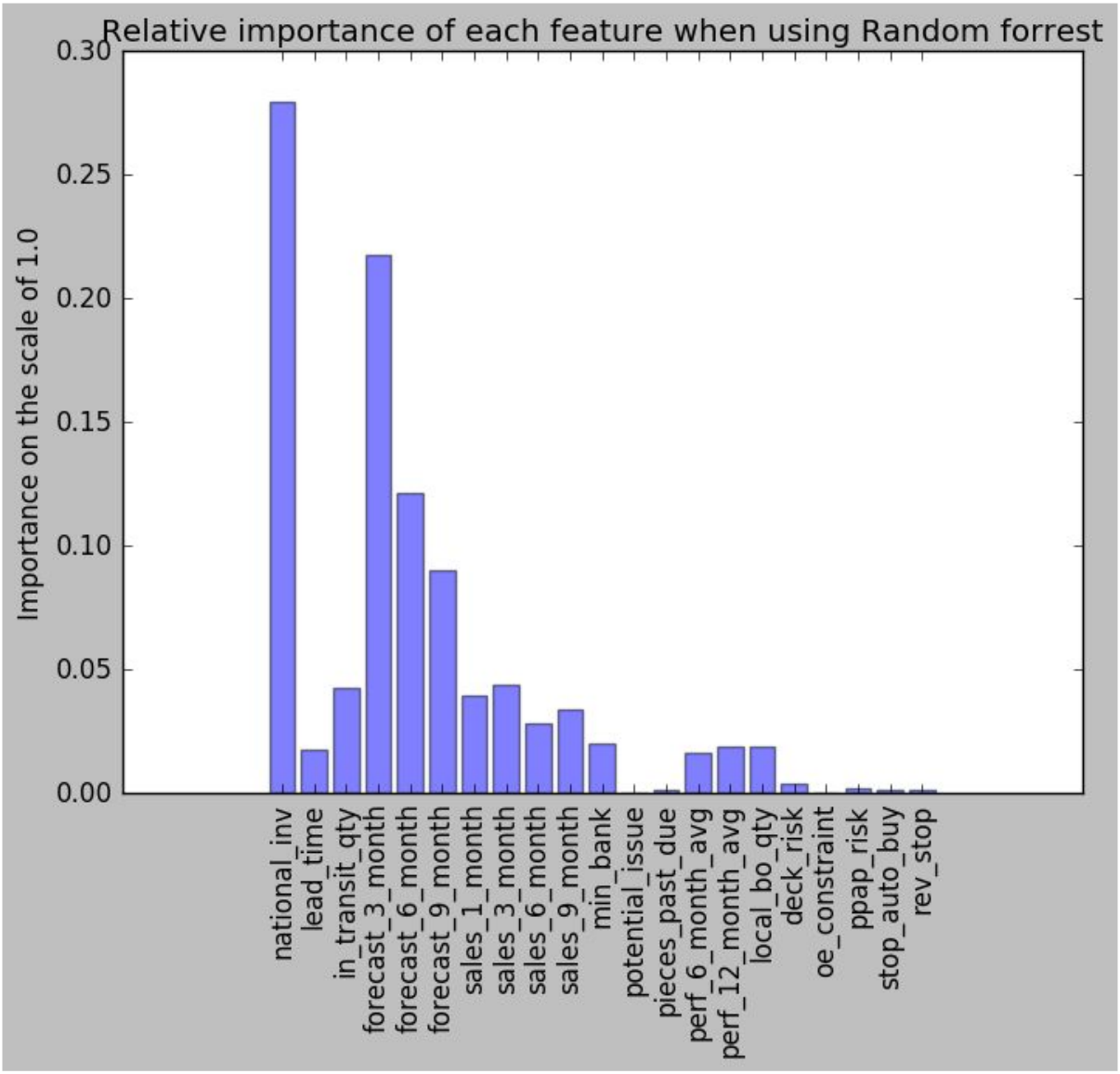
UPDATE(based on review):

- Is the model robust enough for the problem?
    - Yes. It clearly has a high f1 score and precision than our base model.
- Can results found from the model be trusted?
    - We have greatly increased the trust in prediction by 42% better precision and 38% better f1 score. However more work can be done if we get access to more types of data and features.
- Is the final model reasonable and aligning with solution expectations?
    - The expectation we had was to find a substantially better model at prediction. We have achieved this. This prediction can also be augmented with a human who will finally drive the decision. This is helpful because there may be many other datapoints that the human is capable of identifying including both tangible and intangible data.

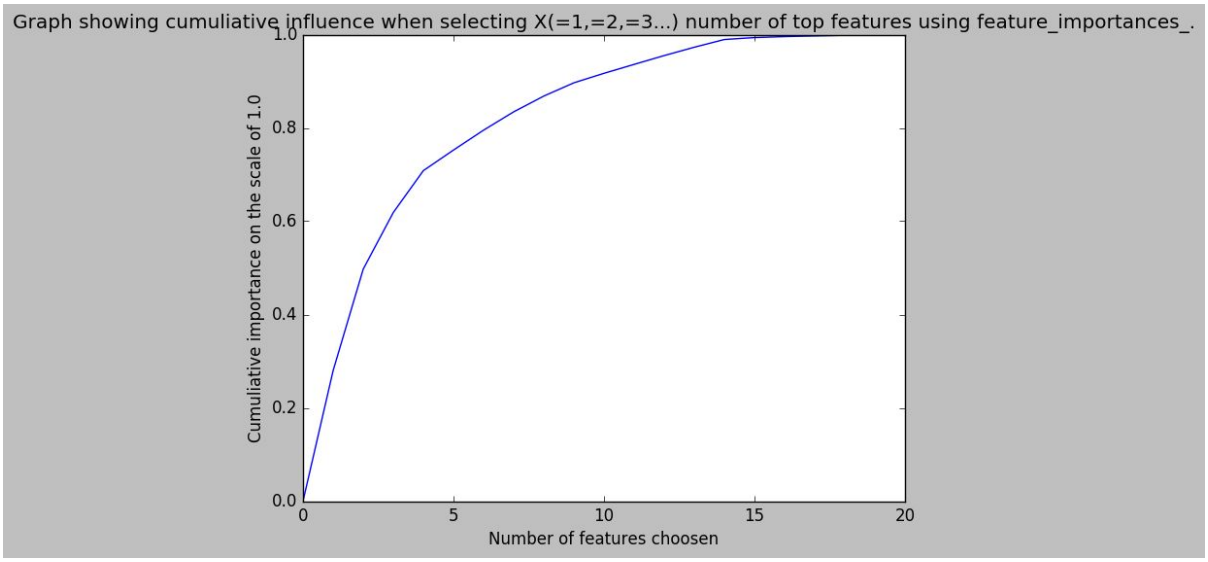
V. Conclusion

Free-Form Visualization

A predictor is useful, but far more useful is intuitive understanding of the problem at hand. It's lucky, then, that the high-performing model of the decision tree can be represented visually. The behavior of the important features will be explored below.



When we produce a graph showing cumulative influences of features, we find that after the first 15 best features the others have very small influences. This helped us explain the parameter that was chosen for max depth



Examining more the feature importances, we see that around 10 features hold most of the importances. Out of this National inventory (national_inv) and forecasting of 1/3/6/9 month sales + actual sales 3/6/9 months are the biggest indicators.

If we examine we can uncover a pattern. If we forecast sales to be high and sales are indeed going strong, the only constraint is the inventory for the part which determines the backorder. No matter what if the inventory low goes down there will be high chances of backorder.

It is also interesting to note that the Boolean columns of features which identified “high risk” actually does NOT yield good correlation. This is a clear non-intutive takeway from this. It is always a combination of current inventory levels + sales levels + forecasting + historical performances that determine the major features.

Using the above we can ask the factory to collect more fine data about these features and experient with our model.

Reflection

The solution to predicting any backorder is to

1. Take in the data for the 15 top features identified
2. Transform any Yes/No values to 1/0 respectively
3. Apply it the previously fit Random forest model with the below parameters
    a. {'criterion': 'entropy', 'max_depth': 15, 'n_estimators': 100}
4. If the prediction is 1, then send a backorder earlier for next week products.

A few interesting stuff I learned are

- We can reduce the feature set be extracting the top features. This not only increases performance due to less dimensionality but also sometimes increases the predictions
- I found that to apply normal SVM on a dataset more than 10k rows its quite difficult and we need to apply LinearSVM which scales better in python
- Neural networks are not well suited in this problem due to very high imbalance of classes. Also the data for backorder true (class=1) is very less
- It took a lot of time to do StatifiedKfold on a Grid search. Due to the many combinations. Thus I had to be more intelligent in choosing which parameter may affect my prediction
- I ran keras on the AWS to make use of the graphic card to run multiple iterations fast and experiment. I learned it in the Udacity self driving Car nano degree and it was useful here

The final solution is a satisfactory solution. This is because business is concerned about 2 things.

1. If they have backorder - this means customer is waiting and its not a good experience.
2. If they send too many inventory into the warehouse, it is wasting space.

Here we have good balance and our solution is better in prediction than benchmark. This can be improved over time after collecting data on how its performing if we fit the model again with new data.

Improvement

1. Our data is very good. I.e it has not many missing peices. However we can design a better system to identify what data should be filled in columns if it is corrupted or not available based on previous data. For example a simple way is to choose the median.
2. We need more data for class=1 where backorder is true. Out data is highly imbalanced and it is a challenge
3. If there are pr-existing solutions we can important from other domain I would try to find out. For example in medical field, we scan the eye to determine eye diseases. Here too the class is very imbalanced between thos who have a particular disease and those who dont. We can research and apply similar techniques to the wearhouse data.

Before submitting, ask yourself. . .

- Does the project report you've written follow a well-organized structure similar to that of the project template?
- Is each section (particularly Analysis and Methodology) written in a clear, concise and specific fashion? Are there any ambiguous terms or phrases that need clarification?
- Would the intended audience of your project be able to understand your analysis, methods, and results?
- Have you properly proof-read your project report to assure there are minimal grammatical and spelling mistakes?
- Are all the resources used for this project correctly cited and referenced?
- Is the code that implements your solution easily readable and properly commented?
- Does the code execute without error and produce results similar to those reported?