DePaul University – College of Computing and Digital Media CSC 478 – Programming Data Mining Applications Professor Bamshad Mobasher

Final Project Report
Project Title: Predicting Cab Booking Cancellations

Student: Naga Venkateshwarlu Yadav Dokku

Spring 2016

Naga Venkateshwarlu Yadav Dokku
CSC-478 Project Report

Abstract:

The business problem tackled here is trying to improve customer service for YourCabs, a cab company in Bangalore. The problem of interest is booking cancellations by the company due to unavailability of a car. The challenge is that cancellations can occur very close to the trip start time, thereby causing passengers inconvenience. In this project I have try to predict possible cancellations of cab booking by the customer using data obtained from the kaggle.com/Competitions. My data analysis model used several methods to analyze the data including Logistic Regression, classification tree, K-nearest neighbor, Gaussian Naïve Bayes and Ensemble algorithms Bagging (Tree,GNB). The accuracy of the model coupled with the final business goal of reducing cost for the company was used to finalize the model for the prediction. The model that selected in the end was Logistic Regression and KNN. Not only does the models have an overall low error rate, but also the cost incurred by the company using this model is the lowest. My recommendation includes running the model in real time on an hourly basis for all pickup times, which are within an hour's time. The model will flag all likely booking cancellations and the operator will call the customers to confirm the booking. Once the operator receives confirmation from the customer, the cab will be dispatched to the pickup location. By using the model for predicting possible customer cancellations, the company will successfully reduce the cost incurred from sending a cab to a pickup location where the customer is not present.

Problem description

Every year the Company, YourCabs loses money due to customer cancellations. The company currently does not have a mechanism to track or predict these cancellations. The company currently only realizes that there is a cancellation when the cab reaches the location; resulting in a cost which can be quantified in such metrics as fuel cost, driver's salary, cab utilization, lost time that the driver which could have been spent attending other bookings and most important lower utilization by the vendors using YourCabs service. This also increases the variable waiting time by the customer. The cost of the cancellations due by customers on a yearly basis is calculated by assuming the average cost of cancellation being Rs. 100 and on average 10% of all booking are cancelled by the customer. This equates to roughly 4, 35,000 (43, 50,000 bookings * 10%*100) in cost each year. My model is meant to predict if the customer will be cancelling his or her booking by not being at the specific location at pickup time

Project Goal

The goal of the project is to create a predictive model for classifying new bookings as to whether they will eventually get cancelled due to car unavailability. This is a classification task that includes misclassification costs. The goal is to find the lowest average-cost-per-booking. Our goal is to reduce the cost incurred by the company as a result of cab cancellations made by the customer. By predicting possible cancellations an hour before the pickup time, YourCabs will be better able to manage its vendors and drivers by providing them with up to date information about customer cancellations and reduce the cost incurred from sending a cab to a booking location that has been cancelled by the customer. Accurate prediction of customer cancellations will lead to a reduction in company costs.

Naga Venkateshwarlu Yadav Dokku
CSC-478 Project Report

Data

For this project, I have used publicly-available datasets from the kaggle.com/Competitions.
Data Source: https://inclass.kaggle.com/c/predicting-cab-booking-cancellations/data

The cab bookings data are made available through a collaboration between Prof. Galit Shmueli at the Indian School of Business and YourCabs co-founder Mr. Rajath Kedilaya and IDRC managing partner, Mr. Amit Batra.
YourCabs is a platform to efficiently connect urban consumers in need of local transport, with vendors in need of increased occupancy.
Industrial Data Research Corp. (IDRC) is a data sciences consultancy focused on Quantitative Modeling, Data Analytics, Scientific Computing, and Data Visualization/Infographics.

Required libraries

This notebook uses several Python packages that come standard with the Anaconda Python distribution. The primary libraries that we'll be using are:

- **NumPy**: Provides a fast numerical array structure and helper functions.
- **pandas**: Provides a DataFrame structure to store data in memory and work with it easily and efficiently.
- **scikit-learn**: The essential Machine Learning package in Python.
- **matplotlib**: Basic plotting library in Python; most other Python plotting libraries are built on top of it.
- **Seaborn**: Advanced statistical plotting library.

Since we're taking into account penalties for misclassification, we can use Weighted Mean Absolute Error

This is the weighted average of absolute errors:

$$\mathrm{WMAE} = \frac{1}{n} \sum_{i=1}^{n} w_i |y_i - \hat{y}_i|$$

Checking the data

The next step is to look at the data.

Generally, we're looking to answer the following questions:

- Is there anything wrong with the data?
- Are there any quirks with the data?
- Do I need to fix or remove any of the data?

Let's start by reading the data into a pandas DataFrame.

Naga Venkateshwarlu Yadav Dokku
CSC-478 Project Report

Data Description

Data fields

- id - booking *ID*
- user_id - the *ID* of the customer (based on mobile number)
- vehicle_model_id - vehicle model type.
- package_id - type of package (1=4hrs & 40kms, 2=8hrs & 80kms, 3=6hrs & 60kms, 4= 10hrs & 100kms, 5=5hrs & 50kms, 6=3hrs & 30kms, 7=12hrs & 120kms)
- travel_type_id - type of travel (1=long distance, 2= point to point, 3= hourly rental).
- from_area_id - unique identifier of area. Applicable only for point-to-point travel and packages
- to_area_id - unique identifier of area. Applicable only for point-to-point travel
- from_city_id - unique identifier of city
- to_city_id - unique identifier of city (only for intercity)
- from_date - time stamp of requested trip start
- to_date - time stamp of trip end
- online_booking - if booking was done on desktop website
- mobile_site_booking - if booking was done on mobile website
- booking_created - time stamp of booking
- from_lat - latitude of from area
- from_long - longitude of from area
- to_lat - latitude of to area
- to_long - longitude of to area
- Car_Cancellation (available only in training data) - whether the booking was cancelled (1) or not (0) due to unavailability of a car.
- Cost_of_error (available only in training data) - the cost incurred if the booking is misclassified. The cost of misclassifying an uncancelled booking as a cancelled booking (cost=1 unit). The cost associated with misclassifying a cancelled booking as uncancelled, This cost is a function of how close the cancellation occurs relative to the trip start time. The closer the trip, the higher the cost. Cancellations occurring less than 15 minutes prior to the trip start incur a fixed penalty of 100 units.

Naga Venkateshwarlu Yadav Dokku
CSC-478 Project Report

Information about the data

	count	mean	std	min	25%	50%	75%	max
id	43431.0	159206.473556	15442.386279	132512.00000	145778.000000	159248.000000	172578.50000	185941.000000
user_id	43431.0	30739.198153	10996.476709	16.00000	24614.000000	31627.000000	39167.00000	48730.000000
vehicle_model_id	43431.0	25.717230	26.798250	1.00000	12.000000	12.000000	24.00000	91.000000
package_id	7550.0	2.030066	1.461756	1.00000	1.000000	2.000000	2.00000	7.000000
travel_type_id	43431.0	2.137252	0.437712	1.00000	2.000000	2.000000	2.00000	3.000000
from_area_id	43343.0	714.544494	419.883553	2.00000	393.000000	590.000000	1089.00000	1403.000000
to_area_id	34293.0	669.490917	400.638225	2.00000	393.000000	541.000000	1054.00000	1403.000000
from_city_id	16345.0	14.915081	1.165306	1.00000	15.000000	15.000000	15.00000	31.000000
to_city_id	1588.0	68.537783	49.880732	4.00000	32.000000	49.000000	108.00000	203.000000
online_booking	43431.0	0.351592	0.477473	0.00000	0.000000	0.000000	1.00000	1.000000
mobile_site_booking	43431.0	0.043241	0.203402	0.00000	0.000000	0.000000	0.00000	1.000000
from_lat	43338.0	12.982461	0.085933	12.77663	12.926450	12.968887	13.00775	13.366072
from_long	43338.0	77.636255	0.059391	77.38693	77.593661	77.635750	77.68890	77.786420
to_lat	34293.0	13.026648	0.113487	12.77663	12.951850	12.982750	13.19956	13.366072
to_long	34293.0	77.640595	0.064045	77.38693	77.582030	77.645030	77.70688	77.786420
Car_Cancellation	43431.0	0.072114	0.258680	0.00000	0.000000	0.000000	0.00000	1.000000
Cost_of_error	43431.0	8.000509	25.350698	0.15000	1.000000	1.000000	1.00000	100.000000

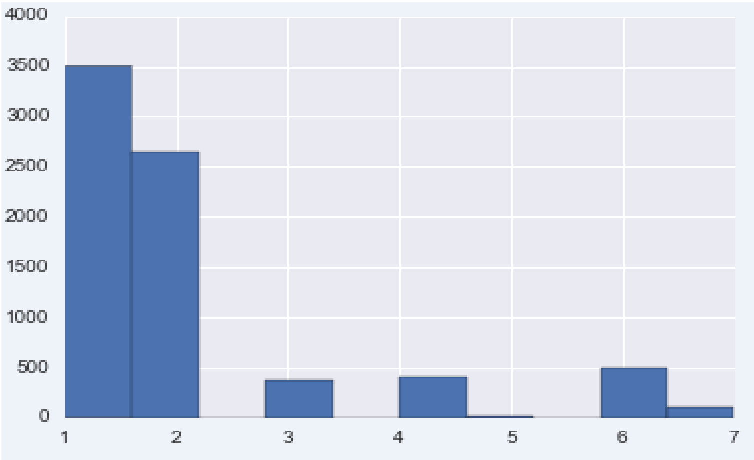
Let’s find out the class balance

```
0     40299
1      3132
```

Classifications (0=no cancellation or 1=cancellation),
Major class imbalance, very few cancellations as compared to large amount of non-cancellations.

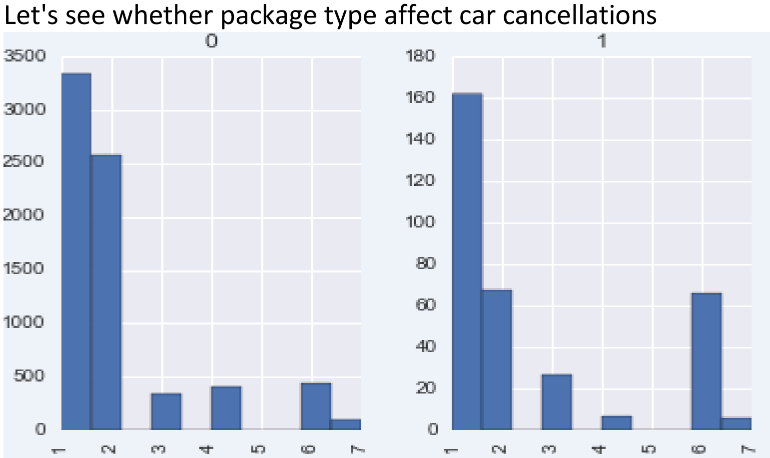
Exploratory-analysis

The distribution of package



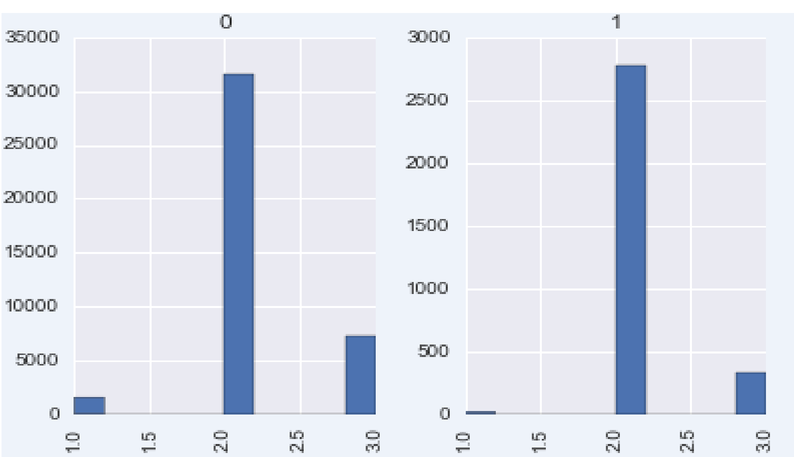
Most of the packages that people opt for are for a journey of 4hrs and around 40kms, followed by 8hrs and 80kms.

Let's see whether package type affect car cancellations



As we can see most of the times 1st package (4hrs & 40kms) gets cancelled followed by packages (3hrs & 30kms) and (8hrs & 80kms).

Let’s take a look at travel_type variable



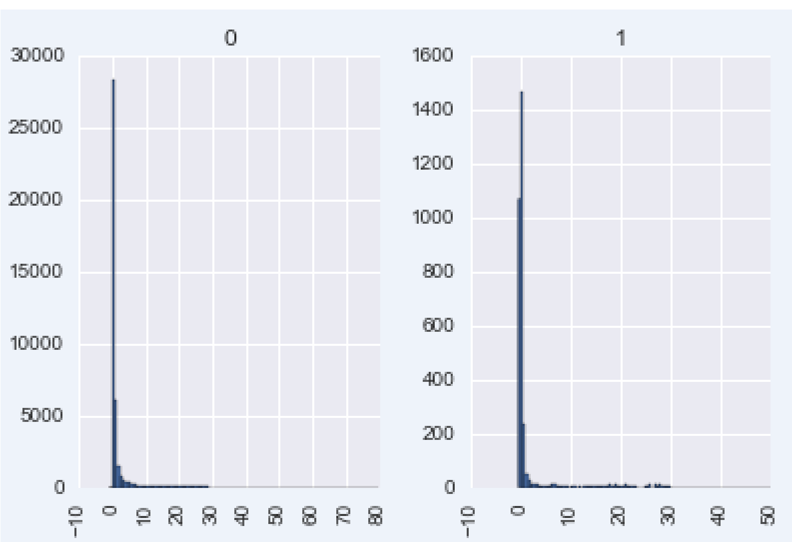
Not surprisingly, most people rent car from point to point travel and around (1 / 10)th of these bookings get cancelled.

130.0	80.000000
1148.0	66.666667
1174.0	66.666667
630.0	66.666667
176.0	52.830189
1381.0	50.000000
1160.0	50.000000
1100.0	50.000000
1385.0	50.000000
1276.0	45.454545
211.0	44.444444
1372.0	40.000000
356.0	40.000000
987.0	40.000000
626.0	34.375000
1258.0	33.333333
34.0	33.333333
326.0	33.333333
177.0	33.333333
833.0	33.333333

So these are areas (from_area) for which more than 50% of the bookings were cancelled.
Lets see if online or mobile booking has any effect on cancellation

online_booking	mobile_site_booking	
0	0	940
	1	289
1	0	1903

And most of the cancellations are of orders that were booked online.



There seems to be no relation between number of days between date of booking and trip's start date with cancellation. Generally people tend to cancel their booking 5 days prior to their trip's start date which is not unusual

Naga Venkateshwarlu Yadav Dokku
CSC-478 Project Report

Methodologies
My goal is to find good algorithm that could reliably possible cancellations of cab booking by the customer. To achieve this, I have tried different models, including Logistic Regression, classification tree, K-nearest neighbor, Gaussian Naïve Bayes and Ensemble algorithms Bagging (Tree,GNB). This problem is a supervised classification problem with all the fields as features and Car_Cancellation as cases.

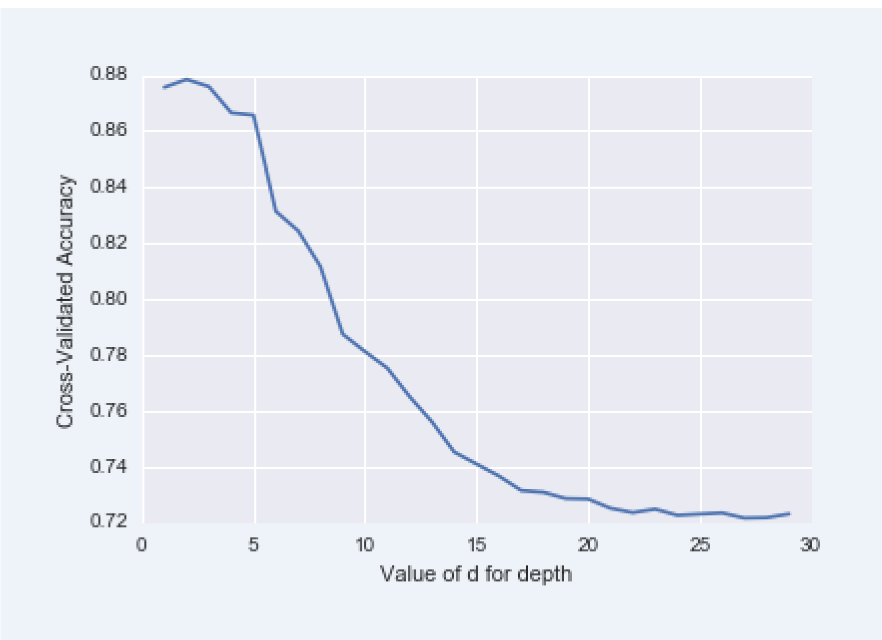
Model Evaluation

10-fold cross-validation Accuracy rate, randomly divide the states into 10 samples and conducted the 10-fold cross-validation. The accuracy rate for all models resulted from the 10-fold cross-validation is shown below.

Model	Null	Gaussian NB	Logistic Regression	Decision Tee	Bagging(Tree)	Bagging(GNB)	Random Forest	K-NN	Voting Classifier
CV Accuracy	0.93	0.87	0.928	0.865	0.863	0.867	0.823	0.9273	0.864

Among the models, Logistic Regression and K-NN generated significantly large accuracy rate comparing to the other models. Null model had the largest accuracy rate among the models but the accuracy rate is similar to that of Logistic Regression and K-NN.

Results:
Model Evaluations without Feature Selection
Decision Tree Classifier
Searched for an optimal value of max_depth for tree classifier



Ensemble Bagging (use Decision Tree as base estimator)

Parameter tuning: n_estimator and max_samples

The plot shows the value of max_samples for bagging (x-axis) versus the cross-validated accuracy (y-axis) each line represents a different value of n_estimators

Ensemble Bagging (use Gaussian NB as base estimator)

Parameter tuning: n_estimator and max_samples

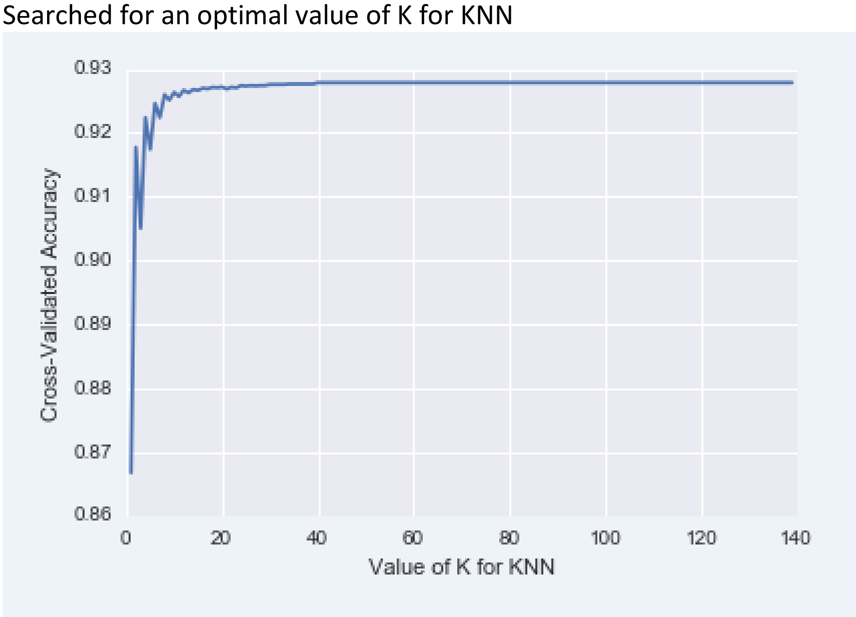
Searched for an optimal values of n_estimator and max_samples for bagging using GNB as the base estimator   the accuracy didn't look right! All of them are about zero???!!!

The plot shows the value of max_samples for bagging (x-axis) versus the cross-validated accuracy (y-axis) each line represents a different value of n_estimators

Naga Venkateshwarlu Yadav Dokku
CSC-478 Project Report

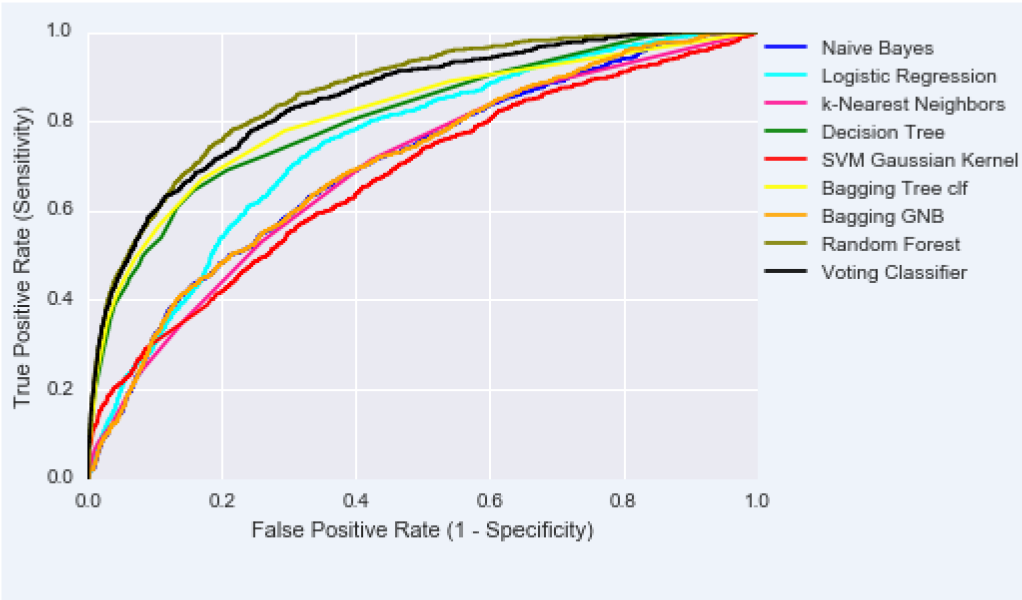
Ensemble Random Forest
Parameter tuning: n_estimator and max_depth
Searched for an optimal values of n_estimator and max_depth for Random Forest Classifier

The plot shows the value of max_samples for bagging (x-axis) versus the cross-validated accuracy (y-axis) each line represents a different value of n_estimators
K Nearest Neighbor
Parameter tuning: k
Searched for an optimal value of K for KNN



The plot shows the value of K for KNN (x-axis) versus the cross-validated accuracy (y-axis

Compare the ROC Curves



Random Forest has the greatest AUC!
SVM Gaussian Kernel showed little prediction power as its ROC curve lies almost on the 45-degree line. Random forest and voting Classifier showed relatively strong prediction power, while Logistic Regression had moderate performance.

Overview of the Model Performance

Model	CV Accuracy	CV AUC	CV Time	AUC	Time
Null	0.927886	0.5		0.5	
Gaussian NB	0.869374	0.738319	0.956	0.705189767546	0.031
Logistic Regression	0.927863	0.759281	20.033	0.749813865133	0.787
Decision Tree	0.865658	0.676056	1.807	0.804150504211	0.069
Bagging (Tree)	0.863355	0.627638	5.924	0.819530195204	0.294
Bagging (GNB)	0.867808	0.738419	5.925	0.705150890245	0.273
Rondom Forest	0.823383	0.541847	6.726	0.858356552084	0.364
Knn	0.927356	0.652422	5.12	0.665510001708	0.441
Voting Classifier	0.864092	0.642437	14.516	0.84666811739	0.865

Of all the procedures and algorithms used, the most useful were cross validation, and using CV AUC and AUC performance metrics.

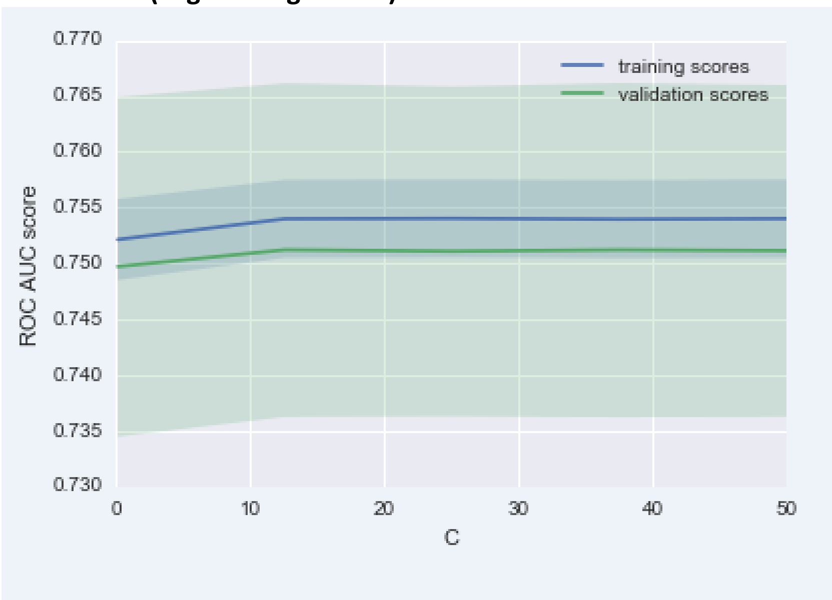
Model Evaluation with feature selection

Feature Importance's Logistic Regression had the best predictability, I used this models to compute the feature Importance's.

Validation curves: plotting scores to evaluate models

It is sometimes helpful to plot the influence of a single hyper parameter on the training score and the validation score to find out whether the estimator is overfitting or underfitting for some hyper parameter values.

Validation (Logistic Regression)



From the above validation curve it is evident that training scores are high than validation scores. So the estimator is overfitting. Since the difference between both curves is not that big, I'm assuming my model does not overfit (a lot).

Learning-Curves (Logistic Regression)

Learning curve shows the validation and training score of an estimator for varying numbers of training samples. To find out how much we benefit from adding more training data and whether the estimator suffers more from a variance error or a bias error. Here both the validation score and the training score converge to a value that is increasing with increasing size of the training set.

Results and Discussion

Besides Random Forest, Bagging (with GNB as base estimator), Decision Tree, Bagging (Tree), Voting Classifier, and all models had accuracies around 0.86 and the null model had high accuracy which was 0.9279.whereas Logistic Regression, Knn did equally better like the null model. As far as the AUC measure, all models had greater AUC than the null model which was 0.5. Random Forest had the next less AUC; Logistic Regression had best predictability, its 10-fold cross-validation accuracies were 0.9278 and AUC was 0.759281.And Random Forest has the greatest AUC! The cross-validation computation of all models except SVM Gaussian Kernel could be finished within a minute. The SVM model took nearly 20 hours to finish the 10-fold cross-validation, but its predictive performance was nearly the worst among all models. I also estimated the importance of each feature using the best models Logistic Regression, Knn. I also tried to remove features with variance lower than 0.8*(1-0.8), but none of the features was lower than this threshold. Therefore, I thought that it would be a good stopping point for my analysis and wouldn't proceed the model evaluations with model selection. From PCA it can be observed that the first 2 components capture (explain) 95% of the variance in the data. We can notice that one vector is longer than the other. In a sense, this tells us that that direction in the data is somehow more "important" than the other direction. The explained variance quantifies this measure of "importance" in direction.

I found through our investigation that Logistic Regression was the best at predicting to whether they will eventually get cancelled due to car unavailability. It produced a prediction with the highest AUC value of 0.759281and 10-fold cross-validation accuracies were 0.9278.
Of all the procedures and algorithms used, the most useful were cross validation, and using CV AUC and AUC performance metrics.

Future Work

Finally, I would want to use libSVM with a nonlinear kernel such as Gaussian to compare with our other algorithms. Due to computational performance limitations, I was unable to implement this method.

References

1. Professor BAMSHAD MOBASHER “CSC 478 - Programming Data Mining Applications”. DePaul University Spring-2016 http://facweb.cs.depaul.edu/mobasher/classes/CSC478/
2. https://inclass.kaggle.com/c/predicting-cab-booking-cancellations/data
3. http://www.scipy-lectures.org/index.html
4. **Machine Learning in Action**, by Peter Harrington, Manning Publications, 2012.