Business Analytics using Data Mining

Project Report

Indian School of Business

Group – A6

Bhushan Khandelwal – 61410182 | 61410806 - Mahabaleshwar Bhat
Mayank Gupta – 61410659 | 61410697 - Shikhar Angra
Sujay Koparde - 61410812

Executive Summary

Idea: Our project involves developing a model for predicting the travel time for a particular cab booking. Such a model holds a lot of value for the consumers as well as the cab company. Many times customers inquire the booking agents about the time it will take to travel from their source to destination but the booking agents are not able to provide a specific time estimates. They only provide a rough estimate based on the distance between the two locations and sometimes this estimate is way off because they do not take into consideration the difference in travel time at different hours of the day and difference in time taken to cover same distance on two different routes.

Value Proposition: Customers plan their travel to ensure they reach the destination on time, therefore they tend to err on the side of caution but sometimes they plan the travel so much in advance that they reach the destination much earlier than required. This is again not a desirous situation. With our model we would be able to predict the travel time quite accurately and this would allow the customers to plan the travel accordingly to ensure they reach the destination in time and not too much in advance.
From Cab Company's perspective such a feature can result in better customer value and higher customer satisfaction levels which could lead to higher customer acquisition and retention levels.

Data: We used the data provided by YourCabs.com for building our model. Each row in the dataset corresponds to a single journey. There were approximately 80,000 bookings in the dataset between November 2011 and November 2013 and the information included locations, time stamps, vehicle id, cancellation status, and more. For our model we needed records with both the start time and end time and thus we could work with only 21,000 records which were for Point to Point travel. We didn't consider intercity travel because of high dependence of intercity travel time on external factors. The main columns of interest for our model were: vehicle_id, from_date & to_date, from_lat, to_lat, from_long, to_long and the output variable timeDiff_inMin. We added some dummy variables for vehicle_id and day and time slot of journey and added a derived column distance based on the "from" and "to" latitude and longitude.

Analytics Solution: Analytics solution for this model involves predicting the travel time for a particular journey. So this is a supervised data mining task. We would have all the predictors (travel day and time slot, vehicle model and "from" and "to" location) at the time of booking and so can predict the time using the model. We could have also made it a classification task rather than a prediction task but that might have involved too many classes to handle. Thus it would be better to predict the time and then add some safety buffer to give the customer a range of travel time rather than a spot prediction.

Recommendations: Such a model/feature obviously needs to be tested thoroughly before going ahead with a full blown rollout. Therefore we recommend rolling out the new feature in a phase wise manner starting with certain routes where travel time can be predicted with high accuracy. The feature should be rolled out in 2-3 popular routes where the number of bookings is significant enough to test the model properly but at the same time it should not be rolled out on the most popular routes because it might impact a large number of customers if there are any issues with the model or implementation. Depending on the success of the new feature and the accuracy of the model the feature can then be rolled out on other routes too.
Secondly to take care of various idiosyncrasies we recommend displaying a time range instead of spot prediction.

Problem Description

This project would address the problem of estimating travel time for a customer. YourCabs would be able to predict the travel time (within a given range), thus giving the user much better travel time predictions and benefit of saving his waiting time.

Business Goal

- **Stakeholder**: Customers and YourCabs team
- **Business Goal Description**: Estimate the time required to travel from point A to point B based on day of the week, car type and time of travel and relevant external data (if available) such as traffic conditions. The project would yield following benefits:
  - For Cab Users: They would be able to estimate their travel time which would allow the users to plan their travel in a better way thus reduce his wait time.
  - For YourCabs Team: A value addition/differentiation service to customers which might help the company to win additional customers.

Another benefit as explained by CEO of YourCabs – this model would also help in estimating the supply of cars by predicting the time a car would take to drop off a customer and be ready for next trip. This is one of the major challenges company is facing as drivers tend to delay the communication.

A time prediction with 80% accuracy would be considered a success. We are keeping a buffer of 20% because of the inherent traffic problems in India such as frequent traffic jams, rallies, etc.

Data Mining Goal

- **Analytics objective:** Predict the estimated travel time from source to destination for a particular journey and time of day
- **Task Description:** This is a supervised, predictive and forward-looking task as the output i.e. future travel time is predicted based on inputs such as source, destination, time of day and the vehicle_model_id chosen
- **Outcome variable:** Continuous variable Time (in minutes)

Data Description

Data is rich, regional dataset on cab bookings facilitated by YourCabs in Bangalore. The dataset captured between November 2011 and November 2013 includes information about over 80,000 cab bookings, their locations, time stamps, id of booking user, vehicle id, cancellation status, and more.

- Each row represents a journey/trip
- Final data used after the preparation *[Annexure 4]*
  1) Complete possible data set: 20995 rows & 7 columns
  2) Top 10 routes: 881 rows & 18 columns

Detailed description of each column is provided in [*Annexure 3*]

Data Preparation

After deciding on the input and output variables the relevant data was prepared for our model. The data was prepared in three parts:

- **Removing the unwanted data:** We observed that many entries in *to_time* column had null values . Since our main prediction relied on *to_time* variable we removed all the null entries which left us

with around 28000 rows. Also when we calculated the time difference (= *to_time* – *from_time*) most of the entries were coming out be *zero*. So we removed such entries as well. We also removed few outliers where time difference was coming out to be very high > 500 min.

- **Keeping/adding the required data:** Since we were predicting the time required for a particular route, we kept only P-2-P data within the city. We assumed that travel to an outside city will include stoppages etc. including such data will not be useful for our analysis. We also added distance column in our data by taking from and end location longitude and latitude values.
- **Creating Data Categories and dummies:** through following visualizations
  - **Bar chart for *from_time [Annexure 1]*:** created 4 categories for *from_time* variable from 7-11 AM, 11-4 PM, 4-10 PM, 10-7 AM. For these four categories we created three dummies *dmorning*, *dafternoon* and *devening*.
  - **Combination chart between to_areaid, from_area_id and count(id) *[Annexure 1]*:** This chart showed us the combination of high frequency routes. So we chose 10 most high frequency routes to these destinations that reduced our data to 800 rows. We finally created 9 dummies for these 10 routes as our input variables.
  - Creating dummy for day of the week & Car types based on knowledge about the car

Data Mining Solution

For predicting the output i.e. expected duration of travel, we used the below predictive models:

1) Multi-linear Regression 2) K-Nearest Neighbor (K-NN)

After the process of data preparation and massaging the data, we partitioned the data into three parts namely training, validation and test in the ratio of 50:30:20. This will help us in building the model on the training data in case of linear regression and training & validation data in case of K-NN and then testing it on test data.

Methods Applied:

1) Multi-linear regression

On fitting the regression model, we get the following details:

The Regression Model

Input variables	Coefficient	Std. Error	p-value	SS
Constant term	27.63209534	2.50517654	0	24986870
dSmallVehicle	-8.19571781	1.9801811	0.00004101	12670.71484
dSedan	-7.69910574	2.1125114	0.00029614	15585.70313
dWeekday	1.21291506	0.88571805	0.17148946	296.6911926
dMorning	32.16956711	1.35973084	0	183611.8125
dAfternoon	23.90462112	1.31217837	0	58820.14063
dEvening	41.53961945	1.15913796	0	565248.5625
Distance	1.77110898	0.05189084	0	833813.3125

Residual df	3658
R-squared	0.389447201
Std. Dev. estimate	26.7534523
Residual SS	2618203

The beta's of the model fitted are in line with the assumptions and the graphs which showed higher concentration of cab booking records during morning and evenings hours, which matches with the peak traffic and demand during these hours. The model also shows a smaller beta value for smaller cars compared to sedans as smaller cars generally take lesser time than sedans for travel within a city like Bangalore. The overall R-squared value of approximately 0.4 is also decent indicating a good fit of the model for the data.

2) K-Nearest Neighbor(K-NN)

We ran the K-NN model for a value of K=10 and the best value of K as suggested by the model was 10.

Performance evaluation:

Choice of measure

We used RMSE (root mean square error) for checking the performance of the models fitted. We observed that both linear regression and K-NN give almost the same values for RMSE. Since both the models show more or less the same performance, we did not consider the ensemble option in this case.

However, for smaller number of records, as in the case of records related to travel to airport and railway station we observed that linear regression model performs better than the K-NN model.

The summary reports of the models fitted are as in the annexure and shown in exhibit 7 and exhibit 8.

The distribution of residuals for linear regression and K-NN are as in exhibits 9 & exhibit 10 respectively.

Benchmarking

We benchmarked our performance measures with the Naïve's rule which for our record set is 75 minutes for a records set of 21674 records i.e. total time/total # of records = 1630691/31674 = 75 minutes

Evaluation

The RMSE value of 25.62 minutes as in exhibit 1 indicates a prediction accuracy of approximately 70% and it is calculated as (RMSE/Naïve rule = 25.62/75 = 0.34 = 66% accuracy)) in case of linear regression and (24.24/75 = 0.32 = 68% accuracy) in case of K-NN.

We also fitted the models for records pertaining to only airport and railway station travels (roughly 7000 records) and we observed that linear regression performed better than K-NN (due to smaller size of record set) with a prediction accuracy of approximately 75%.

On removing the outliers in the dataset and choosing only the top 10 routes as our record set, we observed that our model could predict with an accuracy as high as 85%.

Given that the size of record set can vary depending on the scenario chosen, we chose the multi linear regression model as K-NN model's performance degrades for smaller datasets.

Conclusion

The prediction of travel time would definitely be useful for both the stakeholders namely the company and customers. Following are our recommendations:

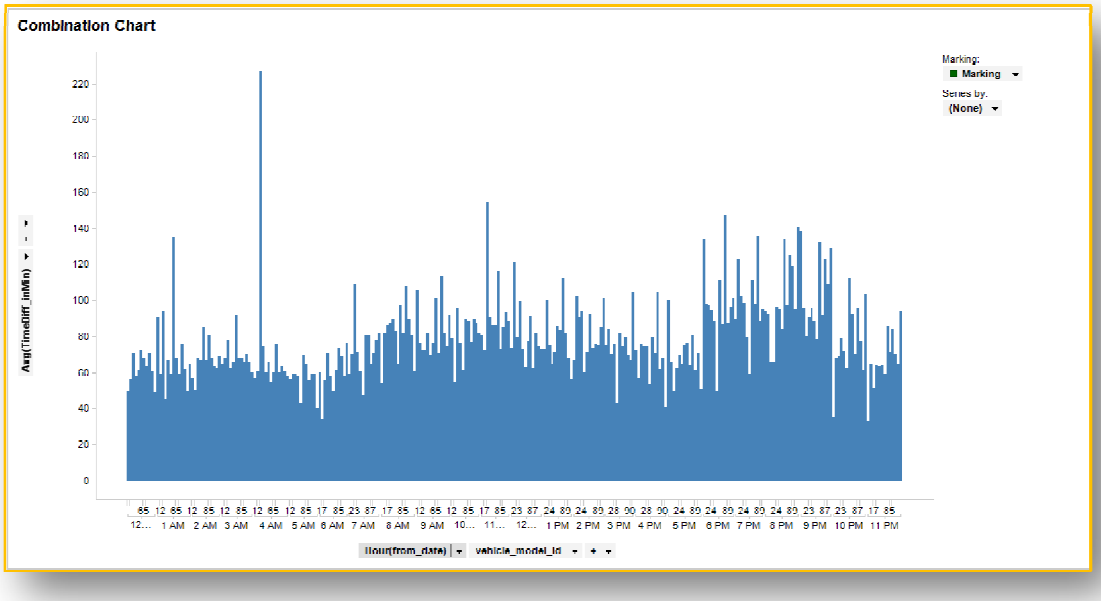
- Roll out the new feature of time prediction in a phased manner starting with certain routes where travel time can be predicted with high accuracy. Initially the feature should not be rolled out on the most popular routes. This will ensure that if there are any specific issues then they do not impact the travel on most popular routes. Instead the feature should be rolled out on routes which lie between 5^{th} and 10^{th} on the popularity ranking. This will allow us to test the feature with substantial amount of travelers without exposing our most popular routes. [***Annexure 6]***. Depending on the success of the new feature it can then be rolled out on other routes too.
- To take care of external factors displaying a time range instead of exact time is recommended.
- Further we also recommend appraising/informing the user that the time prediction does not take into consideration sudden change in traffic conditions due to any external factors such as weather.

Limitations:

- We could not achieve the desired level of accuracy due to outliers. (YourCabs has confirmed that data is erroneous specifically the end time of travel). *[Annexure 5]*
- Model does not consider the traffic conditions and external factors
- Dividing city into zones would probably increase the accuracy

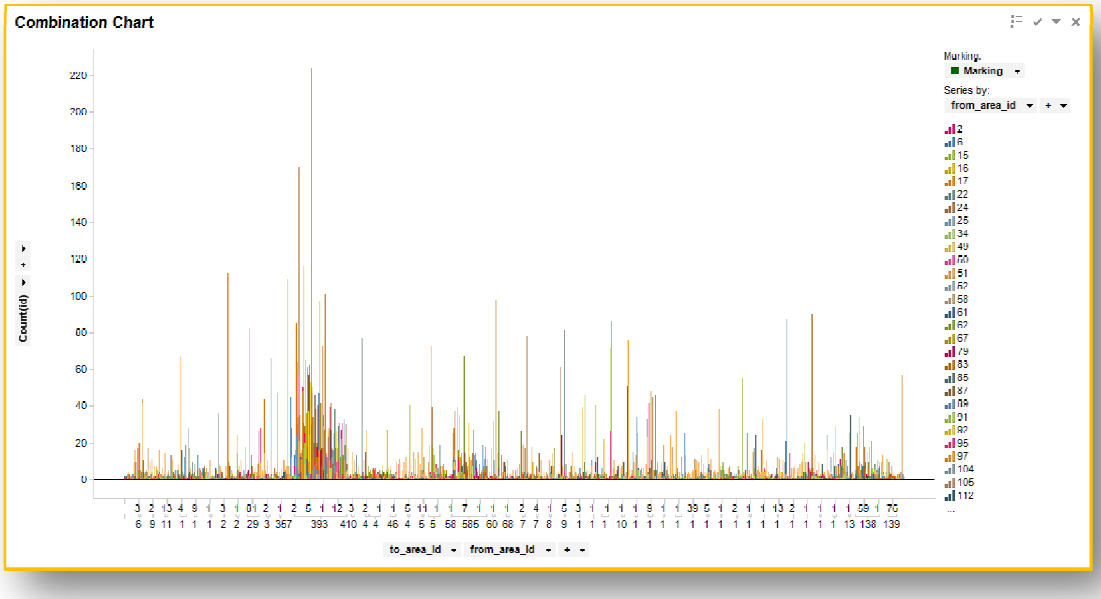
Annexure

1. Demand visualization based on time of the day



Pattern in the above graph was considered to divide 24hours in a day into 4 buckets

Bar chart for *from_time*: We observed that there were two intervals from 7-11 AM and 4-10 PM where there is maximum frequency of travel. Rest 2 intervals from 11AM-4 PM and 10 PM- 7AM didn't attract much travel. So we created 4 categories for *from_time* variable from 7-11 AM, 11-4 PM, 4-10 PM, 10-7 AM. For these four categories we created three dummies *dmorning*, *dafternoon* and *devening*.



2. Finding Busy Routes

Combination chart between to_areaid, from_area_id and count(id):

This chart showed us the combination of high frequency routes. If we see observe the chart, three destinations viz, airport, railway station and MG road Bangalore were attracting maximum amount of travel. So we chose 10 most high frequency routes to these destinations that reduced our data to 800 rows. We finally created 9 dummies for these 10 routes as our input variables.

3. **Data dictionary**

Column Name	Column Description
vehicle_model_id	vehicle model type
travel_type_id	type of travel (1=long distance, 2= point to point (p2p), 3= hourly rental)→We built our model for only Point to Point travel
from_date & to_date	Start and End datetime stamp of journey - ***Not used in the model***
from_lat	latitude of from area - ***Not used in the model***
from_long	longitude of from area - ***Not used in the model***
to_lat	latitude of to area - ***Not used in the model***
to_long	longitude of to area - ***Not used in the model***
dSmallVehicle & dSedan	dummy variables for vehicle_model_id. dSmallVehicle = 1 if small vehicle, dSedan = 1 if sedan and both 0 if SUV/MUV
distance	distance between the start point and end point of journey – calculated using latitude & longitude values /using data from Google Maps
dWeekday	dummy variable for day of the week on the date of travel. 0 if day of travel falls on Saturday or Sunday. 1 otherwise
dMorning	dummy variable for time bucketing. 1 if start time falls between 7 a.m. and 11 a.m.
dAfternoon	dummy variable for time bucketing. 1 if start time falls between 11 a.m. and 4 p.m.
dEvening	dummy variable for time bucketing. 1 if start time falls between 4 p.m. and 10 p.m.
timeDiff_inMin	Output Column – total time taken for the journey. For the existing records (training/validation/test data) this was calculated by taking

the time difference between from_date and to_date in minutes.

---

4. Final data Sample

Considering all the data points

Row Id.	dSmallVehicle	dSedan	dWeekday	dMorning	dAfternoon	dEvening	Distance in km	TimeDiffinMin
1	0	0	0	0	0	0	25.09	123
2	0	1	0	0	0	0	27.97	54
3	0	0	0	1	0	0	27.39	109
4	0	1	1	0	0	1	26.51	68
5	0	1	0	0	1	0	33.59	141
6	1	0	0	0	0	0	26.35	54
7	1	0	0	0	0	0	27.57	68

Considering only top 10 routes

dSmallVehicle	dSedan	dRoute571	dRoute142	dRoute293	dRoute1096	dRoute1010	dRoute1083	dRoute1056	dRoute396	dRoute1065	TimeDiff_inMin	dWeekday	dMorning	dAfternoon	dEvening	Distance
0	1	1	0	0	0	0	0	0	0	0	73	0	0	0	0	48.8
1	0	1	0	0	0	0	0	0	0	0	68	0	0	0	0	48.8
0	1	0	1	0	0	0	0	0	0	0	63	0	0	0	0	42.9
1	0	1	0	0	0	0	0	0	0	0	58	0	0	1	0	48.8
0	0	0	0	1	0	0	0	0	0	0	79	0	1	0	0	52.7
0	1	0	1	0	0	0	0	0	0	0	63	0	0	1	0	42.9
1	0	1	0	0	0	0	0	0	0	0	58	0	1	0	0	48.8
1	0	1	0	0	0	0	0	0	0	0	58	0	1	0	0	48.8
0	1	0	0	0	0	0	0	0	1	0	61	0	0	0	0	50
1	0	0	0	0	0	1	0	0	0	0	64	0	0	1	0	51.6

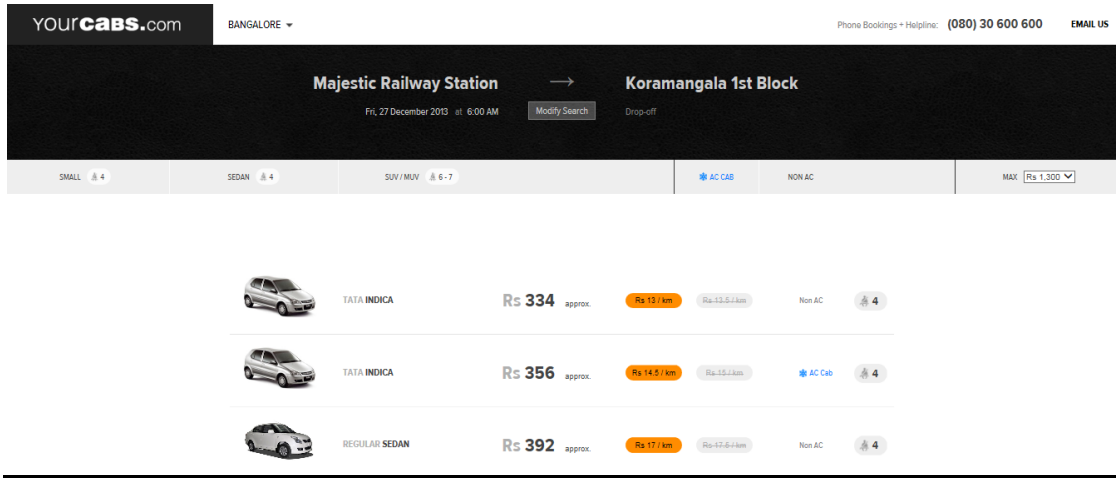
5. Erroneous data sample

Predicted Value	Actual Value	Residual	dSmallVehicle	dSedan	dWeekday	dMorning	dAfternoon	dEvening	Distance
89.376999	21	-68.376999	0	1	0	0	0	0	24.377842
89.839048	17	-72.839048	1	0	1	0	0	0	32.53963
88.899518	149	60.100482						0	7.9122705
88.704777	133	44.295223						0	10.169565
85.57141	139	53.42859						1	1.5644999
98.182346	149	50.817654						0	12.797752
85.397725	142	56.602275						1	4.2370328
87.876512	39	-48.876512	1	0	0	0	1	0	39.976991
87.536647	16	-71.536647	1	0	0	0	1	0	33.274181
65.616117	94	28.383883	0	1	0	0	0	0	2.7318935

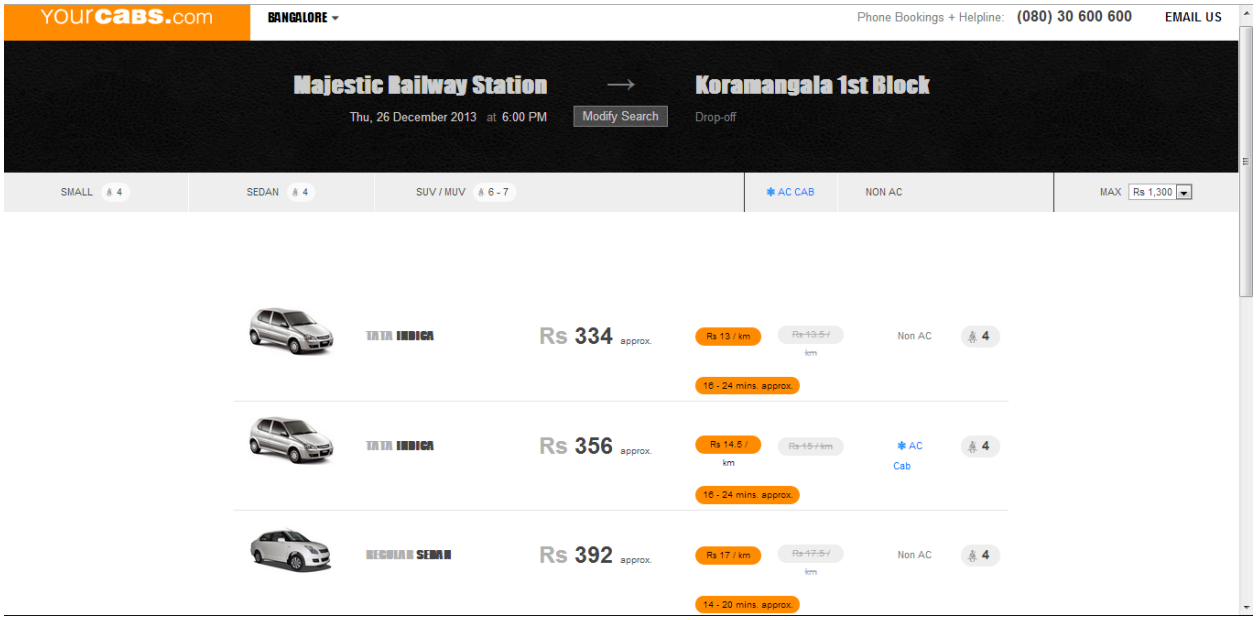
Time required is less than the total distance, this is practically impossible in Bangalore

6. **Sample Concept of usage on YourCabs.com webpage**

Before:



After:



7.

Training Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error

2618203.149	26.72424473	1.31665E-06

Validation Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
1311029.462	24.41707353	-0.12607358

Test Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
962957.7471	25.62929327	-0.53262161

8.

Training Data scoring - Summary Report (for k=10)

Total sum of squared errors	RMS Error	Average Error
1274032.635	11.28730541	0.001412209

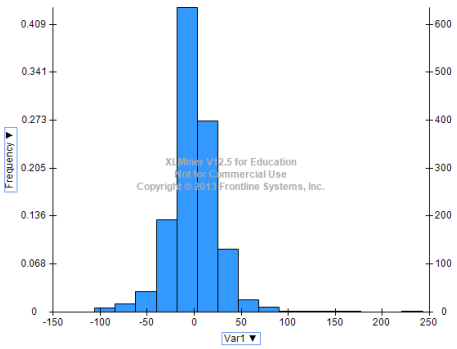
Validation Data scoring - Summary Report (for k=10)

Total sum of squared errors	RMS Error	Average Error
3972175.385	24.54182235	-0.12621997

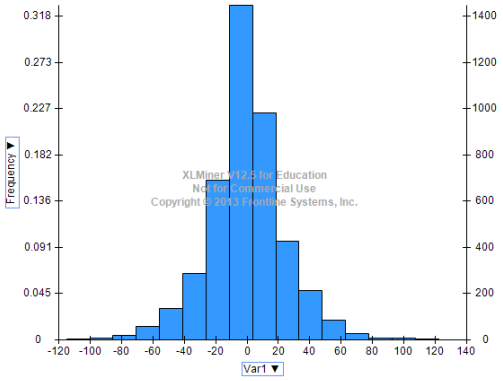
Test Data scoring - Summary Report (for k=10)

Total sum of squared errors	RMS Error	Average Error
2584652.566	24.24779459	-0.1369072

9.



10.


0.318
0.273
0.227
0.182
0.136
0.091
0.045
0
1400
1200
1000
800
600
400
200
0
Frequency
-120 -100 -80 -60 -40 -20 0 20 40 60 80 100 120 140
Var1
XLMiner V12.5 for Education
Not for Commercial Use
Copyright © 2013 Frontline Systems, Inc.