Goal:

The goal is to train model on the history of the ship/vessel's positions and predict the latitude and longitude of a vessel for any given time in the future.

Data Collection:

Using website: marinecdre.gov

- Data for this project has been fetched from the link (https://marinecadastre.gov/ais/)
- The data contains records for U.S. coastal waters for calendar years 2009, 2010, 2011, 2012, 2013, and 2014. Records are filtered to one minute and formatted in zipped, monthly file geodatabases (in .gdb format) by Universal Transverse Mercator (UTM) zone.

Data Filtering:

The main concern here is the size of data and how to shortlist the chunk of such a massive data to train the model and make the best predictions.

Rough Estimate of data size:

Number of Years	Number of months	Number of zones	Approx. min size	Approx. max size	Approx. Average size
1	1	1	1.2 MB	2.5 GB	1.2506GB
1	1	20	20*1.2 = 24MB	20*2.5 = 50GB	25.012GB
1	12	20	12*24 = 288MB	12*50 = 600GB	300.144GB
6	12	20	6*288 = 1.73GB	6*600 = 3600GB	1800.865GB

Conversion of data from .gdb format to .csv format:

To parse the data, we needed special tools that are either proprietary or need special permissions, like

- To convert the gdb to a csv we required paid proprietary software called ARCGIS (the student version is $100) which provides a "arcpy" python library to parse the gdb data.
- We were able to parse the .gdbtable files using an open source library but the remaining files (.gdbindexes, ,gdbtablx etc.) failed.
- QGIS, GDAL and OSGI were other tools that were also looked into but with little success.
- Finally, we found the open source python library 'fiona' which helped us in parsing the data and converting the gdb files format to .csv format as follows:

	latitude	longitude	SOG	COG	Heading	ROT	Timestamp	Status	VoyageID	MMSI	ReceiverType	ReceiverID
0	52.932243	-176.887782	19.299999	275.600006	277.0	7.0	1325742895	0	1	636190187	r	17MADA1
1	52.932870	-176.897497	19.299999	276.200012	277.0	246.0	1325742959	0	1	636190187	r	17MADA1
2	52.933532	-176.908863	19.200001	274.500000	277.0	3.0	1325743038	0	1	636190187	r	17MADA1
3	52.934185	-176.918492	19.299999	276.200012	277.0	8.0	1325743100	0	1	636190187	r	17MADA1
4	52.934913	-176.929183	19.299999	276.500000	277.0	251.0	1325743174	0	1	636190187	r	17MADA1
5	52.935637	-176.940637	19.299999	275.700012	276.0	11.0	1325743253	0	1	636190187	r	17MADA1
6	52.936472	-176.954870	19.299999	275.500000	277.0	0.0	1325743350	0	1	636190187	r	17MADA1
7	52.940480	-177.017113	19.200001	275.399994	277.0	251.0	1325743775	0	1	636190187	r	17MADA1
8	52.950698	-177.180103	19.299999	275.899994	277.0	13.0	1325744883	0	1	636190187	r	17MADA1
9	52.953170	-177.216898	19.200001	275.200012	277.0	17.0	1325745137	0	1	636190187	r	17MADA1
10	52.955283	-177.249445	19.299999	275.700012	276.0	7.0	1325745298	0	1	636190187	r	17MADA1
11	52.956183	-177.263347	19.200001	275.200012	277.0	250.0	1325745455	0	1	636190187	r	17MADA1
12	52.956837	-177.273145	19.200001	275.600006	276.0	10.0	1325745523	0	1	636190187	r	17MADA1
13	52.957460	-177.282790	19.200001	276.399994	277.0	244.0	1325745587	0	1	636190187	r	17MADA1

CSV contains following 15 attributes:
Latitude: Latitude
Longitude: Longitude
SOG: Speed over Ground
COG: Course Over Ground
Heading: True Heading
ROT: Rate Of Turn
Timestamp: Timestamp
Status: Code for navigational status
VoyageID: Unique Identifier for each vessel
MMSI: Encoded Maritime Mobile Service Identity; Unique identifier for ship stations, ship earth stations, coast stations, coast earth stations, and group calls
ReceiverType: Type of message receiving unit: Base Station (b) or Receiver ®
ReceiverID: ID number of the receiver

Column filtering:

1. The columns **ReceiverID, ReceiverType, MMSI** represented either IDs or encoded data. These columns are not going to show massive improvement in the model. Hence, We dropped these columns.
2. Timestamp being a long integer might not make the model great.So, we decided to split timestamp into Year, Month, Day, Hour, Minute, Second.
3. As we are considering first 3 months of 2014 for the prediction, dropping the column **Year** makes sense.

The final csv looks like as follows:

	latitude	longitude	SOG	COG	Heading	ROT	Status	VoyageID	Month	Day	Hour	Min	Second
0	53.320122	-176.940237	9.700000	80.099998	92.0	0.0	0	11	1	3	13	35	44
1	53.321838	-176.925508	9.600000	77.300003	87.0	127.0	0	11	1	3	13	39	3
2	53.326722	-176.883112	10.300000	78.500000	90.0	0.0	0	11	1	3	13	48	34
3	53.334162	-176.818292	9.700000	82.900002	91.0	0.0	0	11	1	3	14	2	55
4	-83.967050	-175.831042	38.900002	388.700012	473.0	0.0	0	11	2	1	20	36	45
5	51.265167	-178.160050	11.800000	72.800003	72.0	127.0	0	11	3	2	6	20	56
6	51.268845	-178.139318	12.100000	71.699997	71.0	129.0	0	11	3	2	6	24	56
7	51.270350	-178.130733	12.100000	75.099998	73.0	0.0	0	11	3	2	6	26	36
8	51.273565	-178.111143	13.000000	77.599998	77.0	127.0	0	11	3	2	6	30	16
9	51.278583	-178.084667	12.500000	68.599998	69.0	129.0	0	11	3	2	6	35	17
10	51.280133	-178.076717	12.600000	72.099998	72.0	129.0	0	11	3	2	6	36	46
11	51.282633	-178.064867	11.500000	72.500000	72.0	127.0	0	11	3	2	6	39	6
12	51.283617	-178.059767	12.300000	73.900002	73.0	0.0	0	11	3	2	6	40	5
13	51.284667	-178.054700	11.900000	70.599998	70.0	129.0	0	11	3	2	6	41	6
14	51.285717	-178.049500	12.600000	72.400002	72.0	0.0	0	11	3	2	6	42	6
15	51.287267	-178.041383	12.600000	72.699997	72.0	0.0	0	11	3	2	6	43	35
16	51.289390	-178.030415	13.100000	73.500000	73.0	0.0	0	11	3	2	6	45	36
17	51.292188	-178.016102	11.900000	71.300003	71.0	129.0	0	11	3	2	6	48	17
18	51.293363	-178.009995	12.100000	73.000000	73.0	127.0	0	11	3	2	6	49	27
19	51.294517	-178.003900	12.700000	72.500000	72.0	0.0	0	11	3	2	6	50	36

Outlier Removal in the dataset with ~22 k vessels , zone 11 and 12 months of year 2012:

We found outliers during the plot of longitude (x - axis) and latitude (y-axis) for a voyage Id. These Outliers were removed by fixing the range of longitude, latitude for that particular zone and chopping off all the rows out of the range. Below plots show the outliers for two vessels :

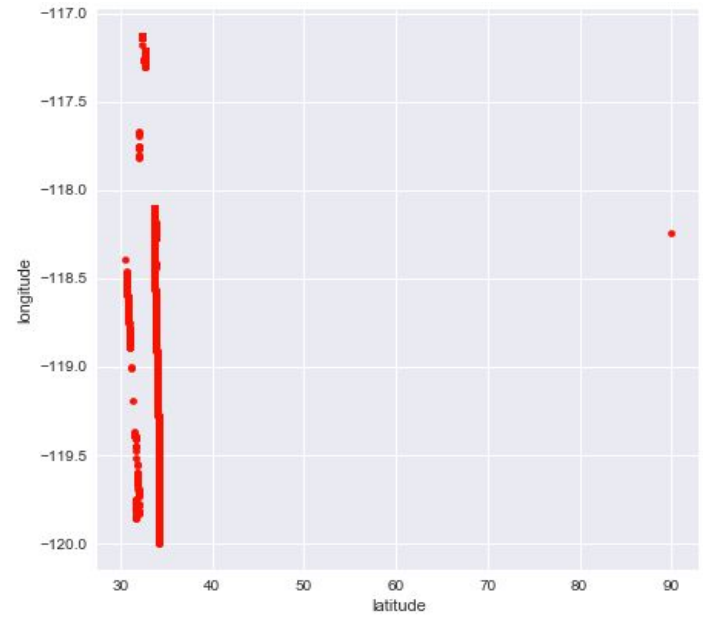


fig showing outlier in case of voyage id = 228

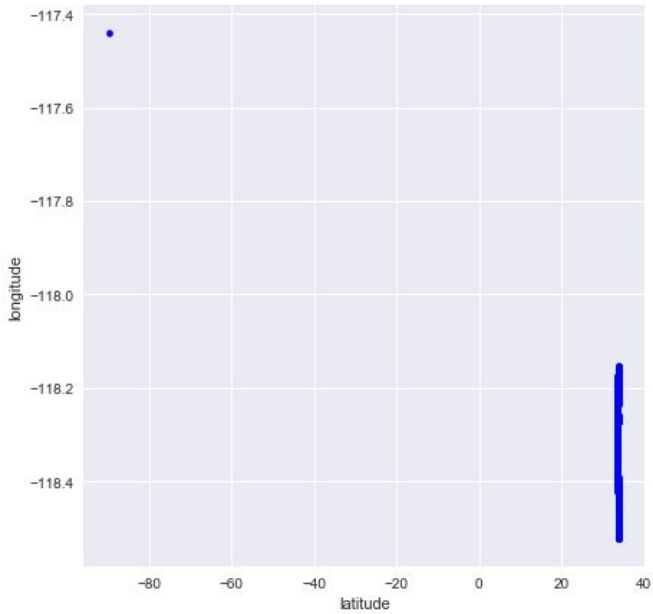


fig showing outlier in case of voyage id = 6136

Linear Regression Model :

Training and prediction of model in phase 1	**Training and prediction of model in this phase**
Training the model on first 3 months of a particular year	Training the model on first n months of a particular year
Predicting the location of the ship on the 4th month of the same year	Predicting the location of the ship on the remaining months of the same year

We have trained the model on first *8* months (it can be any number, we chose this because beyond this ,our machines got crashed) of *2012* and predicted the longitude and latitude of remaining *4* months of the *2012*.

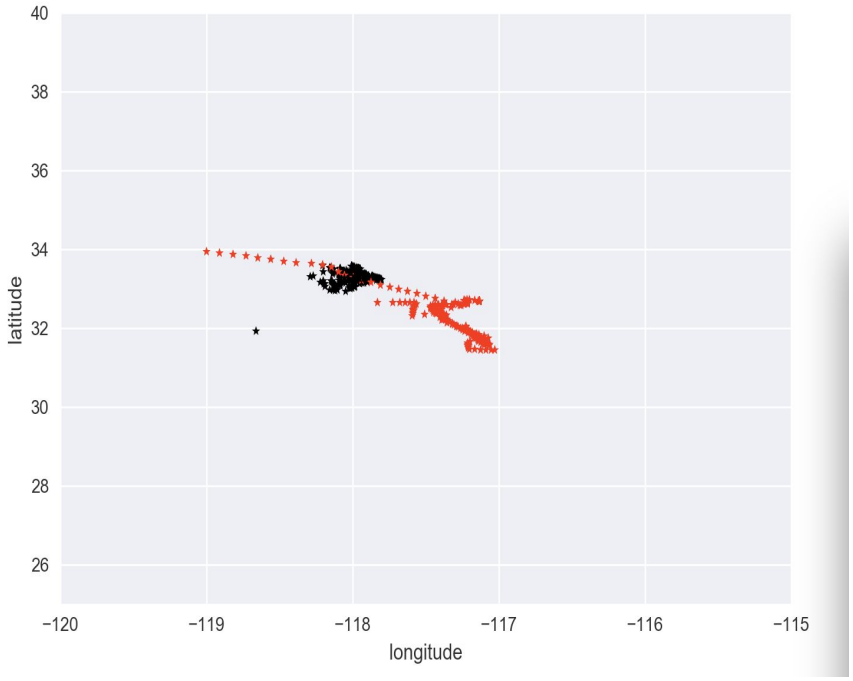
With this dataset, We got our linear regression model improved. Mean square error for latitude reduced by ~ *160* times and mean square error for longitude reduced by ~ *426* times. (0.40836 for Latitude and 0.30307 for Longitude)

Plotting the actual (Marked in red) and predicted (Marked in black) values of longitude and latitude :

We calculated the average SOG for all the vessels present in testing dataset. This helped us to analyse that there are many ships with no considerable movement or zero movement With this, we came up with roughly three classes of ships on the basis of average SOG:

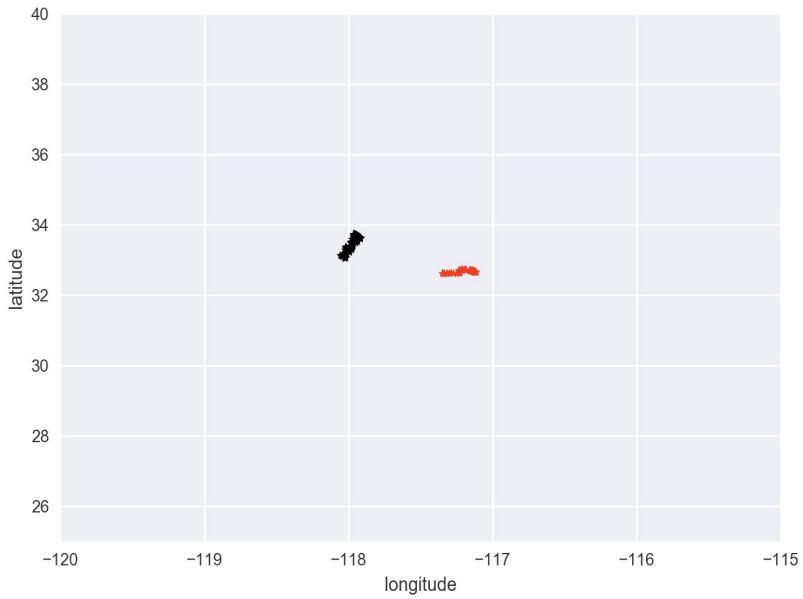
- Fast moving ships (SOG ranging ~ between 53 knot to 103 knot)
- Average moving ships (SOG ranging ~ between 3 knot to 53 knot)
- Slow moving ships (SOG ranging ~ between 0 knot to 3 knot)

Plotting the actual and predicted trajectory of ship (Class : fast moving ship) with increasing timestamp (Red represents the actual data and black represents the predicted data) :
Voyage Id = 2353
Average SOG = 102.300003 knot



Plotting the actual and predicted trajectory of ship (Class : average moving ship) with increasing timestamp (Red represents the actual data and black represents the predicted data) :
Voyage Id = 95
Average SOG = 20.384259 knot

Plotting the actual and predicted trajectory of docked ship (Class : slow moving ship) with increasing timestamp (Red represents the actual data and black represents the predicted data) :
Voyage Id = 354
Average SOG = 1.426 knot



With the same dataset, We explored XGB model, which couldn't outperform the simpler Linear Regression Model. Mean square error for latitude reduced was 0.51 and mean square error for longitude was 0.6

Predicting the Speed over Ground (SOG) instead of Location (longitude and latitude) attributes:

We also considered an alternate baseline model, that predicts the Speed over Ground instead of two separate models for predicting latitude and longitude.

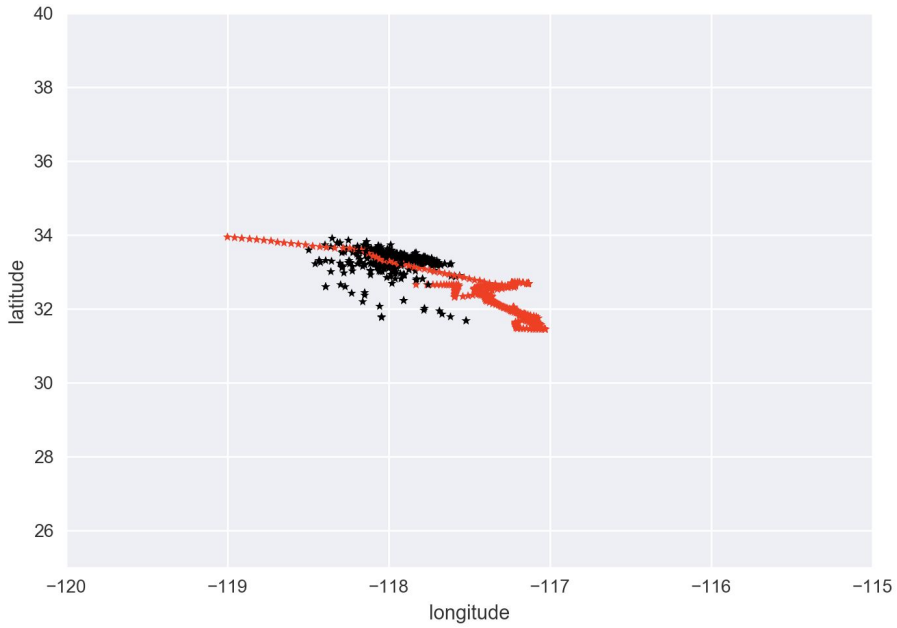
The reasoning for the above consideration is given the current latitude and longitude, if we have the heading (direction) information and the predicted speed, we can easily compute the distance by multiplying the speed with the time difference (future and current time) and predict the latitude and longitude by extrapolation in the given direction.

The true heading attribute plays a crucial role as it gives us the required direction in terms of angle made with respect to the "True North Pole".

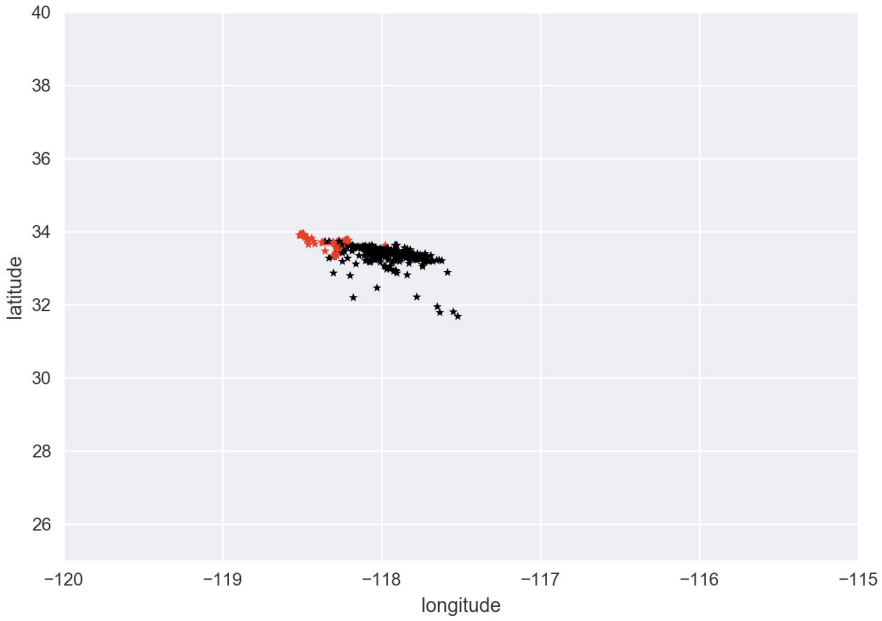
The modeling was based on the a similar approach like the latitude and longitude. These attributes discussed so far are very likely to have a linear correlation with the speed/latitude/longitude and hence we opted for Linear Regression Models.

Plotting the actual and predicted trajectory of ship (Class : fast moving ship) with increasing timestamp (Red represents the actual data and black represents the predicted data) :

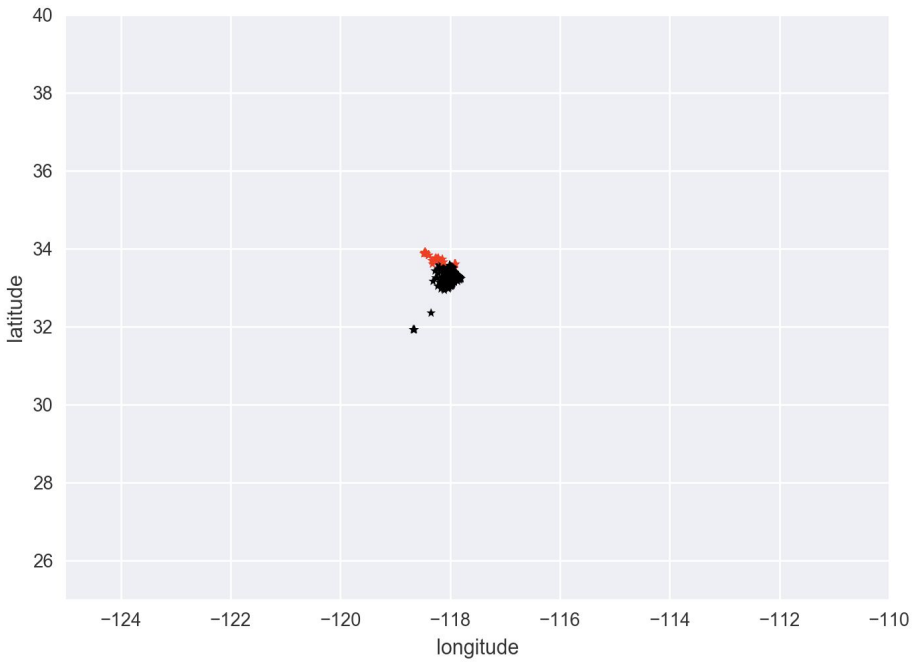
Voyage Id = 2353



Plotting the actual and predicted trajectory of ship (Class : average moving ship) with increasing timestamp (Red represents the actual data and black represents the predicted data) :
Voyage Id = 95



Plotting the actual and predicted trajectory of docked ship (Class : slow moving ship) with increasing timestamp (Red represents the actual data and black represents the predicted data) :
Voyage Id = 354



Improvement in the prediction results after the inclusion of an external dataset:

External Data Collection:

We had decided to include the following as part of the external datasets:

1) Congestion at ports
2) Weather
3) Trade Embargos
4) Cargo type

Since the dataset didn't disclose any information about the type of vessel, the effect of "cargo" on the location prediction could not be explored. The ship's Speed over Ground attribute carries this information to some extent.

There wasn't any "trade embargo" for the UTM zone taken for modeling and neither the congestion at ports would have been part of the existing dataset but the interconnection between any two ships couldn't be studied due to confidentiality of the data.

Historical Marine Weather Dataset:

This dataset consisted of end to end information for any given latitude and longitude. The data of interest was only the points in the ocean and most of the freely available datasets were open for developers to access locations on land.

Most marine websites provided free access to marine data in real-time and our point of interest lies in 2012 data. The following portals allowed developers to access historical data using a valid registered key and Restful APIs:

1) https://developer.worldweatheronline.com/
2) https://www.wunderground.com/weather/api/d/docs
3) http://www.metocean.co.nz/apis/
4) https://www.worldtides.info/

Historical weather data from worldweatheronline is in Beta testing stage and hence not available for immediate usage.
The data from wunderground was fully paid and that of metocean was also restricted by sales team for any free access requirement.

Factors such as temperature and humidity was unlikely to affect the course of the ship's journey. Tidal information would be a significant improvement if obtained and processed in the right way.

Worldtides website provided information about the tidal waves at any given point in ocean. 1000 credits per month was also provided free of cost to use the API. Each API call that was made had the following attributes sent,

1) Latitude
2) Longitude
3) Timestamp
4) Length of days to fetch the data
5) Maximum Number of internal API calls
6) Private Key

And returned a JSON object with the following information

1) Timestamps for information recorded at the given position
2) Height of tidal waves from the sea level for each of the days spanning across every 30 mins if available

The most important input to obtain the external data set were latitude longitude and the time. The data for modeling was chosen for the particular zone ranging from -90 degree to +90 degree in latitude and -120 degree to -114 degree in longitude. The time had to be taken for all the 365 days.

The API allowed querying for multiple days' worth of data at the cost of extra credits. A total of 3000 credits (3 x 1000) had to be divided in the following manner.

1) Divide the zone into smaller groups with dimensions of 15 degree latitude and 3 degree longitude
2) Send the center of each of those groups to get the tidal data.
3) The length of the days in query was set to 14 days

Total number of API calls would be = (365/14) * (180 * 6) / (15 * 3) ~= 625 calls
Total number of credits used = 625 * 4 = 2500 credits

Each of this group roughly represents 180,000 square miles approximately. This could have been reduced further to improve the accuracy of tidal waves but the constraints mentioned above prevented us from doing so.

On finding the center points of the above mentioned groups, 7 out of 24 were on land (Locations on Antarctica, USA and Canada). To overcome this drawback, we had to manually remove those points and replace them with closest and perpendicular points in the ocean (not close to the coast, as the ships are usually docked few miles away from it).

Invalid co-ordinates -82.5, -118.5 --> Valid Co-ordinates -70, -138
Invalid co-ordinates -82.5, -115.5 --> Valid Co-ordinates -70, -138
Invalid co-ordinates 37.5, -118.5 --> Valid Co-ordinates 37.5, -125

Invalid co-ordinates 37.5, -115.5 --> Valid Co-ordinates 37.5, -125
Invalid co-ordinates 52.5, -118.5 --> Valid Co-ordinates 52.5, -135
Invalid co-ordinates 52.5, -115.5 --> Valid Co-ordinates 52.5, -135
Invalid co-ordinates 67.5, -118.5 --> Valid Co-ordinates 75, -135

External Data Parsing:

The final external dataset consisted of a dictionary/JSON where each latitude range midpoint was mapped to each longitude range midpoint and each of those were further mapped to the month, day and hour in a nested manner.

Latitude -> Longitude -> Month -> Day -> Hour

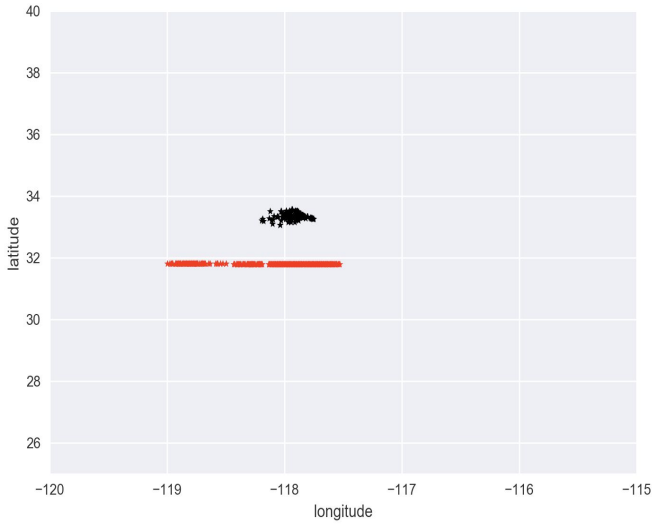
On execution of the above process, we had tidal height information for each of 365 days in the year 2012. There were missing information in this tidal dataset where few hours of tidal information was missing. To accommodate this, we replaced the closest non-zero tidal value while populating the original information by efficiently looking at the left and right halves of the missing/zero values.

The entire 2012 data set (consisting of almost 50 million rows of almost 5GB data) was iterated over and the tidal information from the external dataset was populated for modeling.

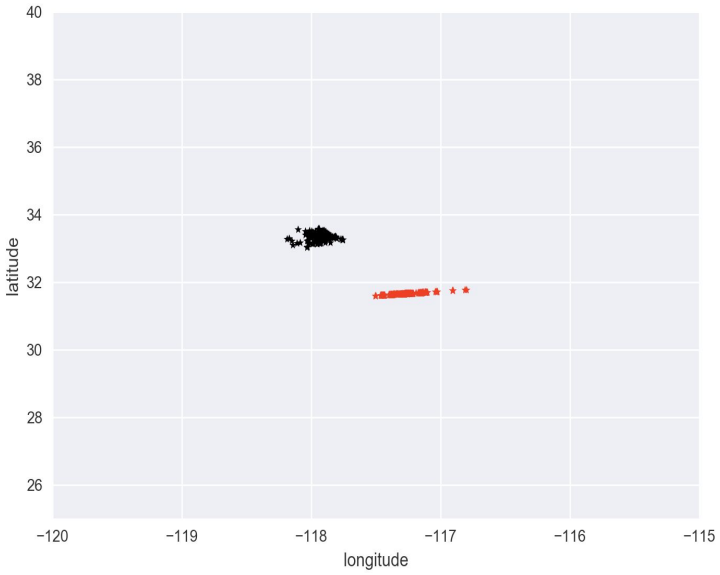
Linear Regression model after adding the new external data set :

Plotting the actual and predicted trajectory of ship (Class : fast moving ship) with increasing timestamp (Red represents the actual data and black represents the predicted data) :

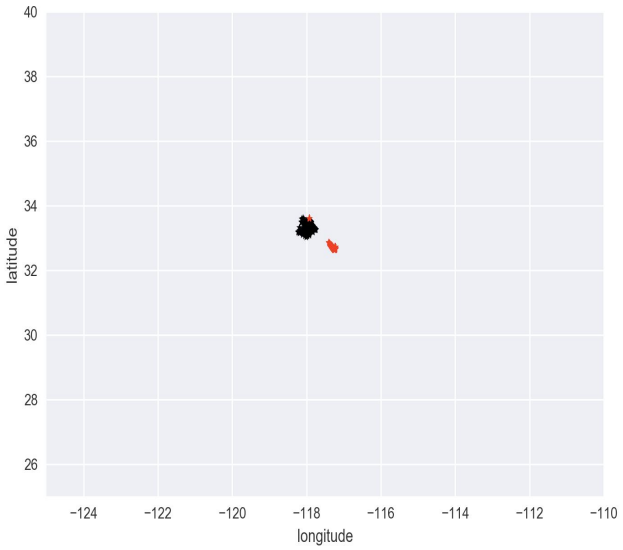
Voyage Id = 2823
Average SOG = 85.7067 knot



Plotting the actual and predicted trajectory of ship (Class : average moving ship) with increasing timestamp (Red represents the actual data and black represents the predicted data) :
Voyage Id = 82
Average SOG = 19.266517 knot



Plotting the actual and predicted trajectory of docked ship (Class : slow moving ship) with increasing timestamp (Red represents the actual data and black represents the predicted data) :
Voyage Id = 115
Average SOG = 1.904215


40
38
36
34
32
30
28
26
latitude
−124
−122
−120
−118
−116
−114
−112
−110
longitude