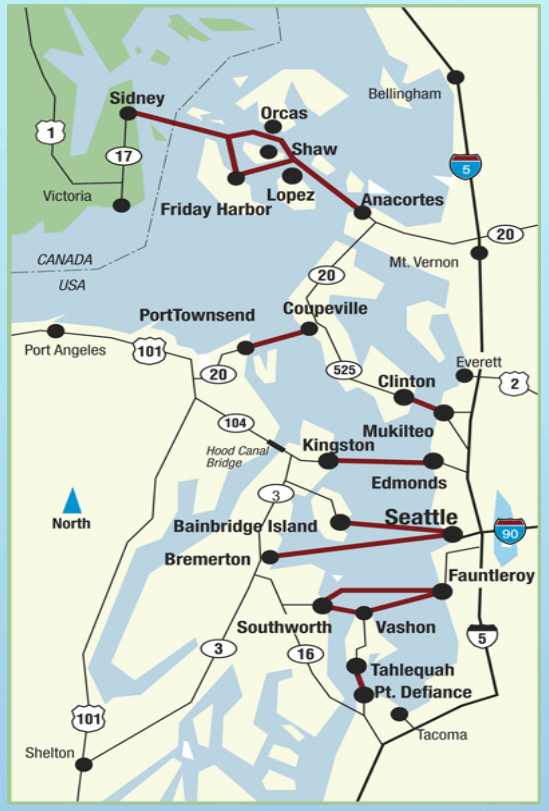
Prediction of Ferry Arrival Times Using Machine Learning

Mehmet Duman
University of Massachusetts, Dartmouth

Faculty Advisor
Dr. David Koop

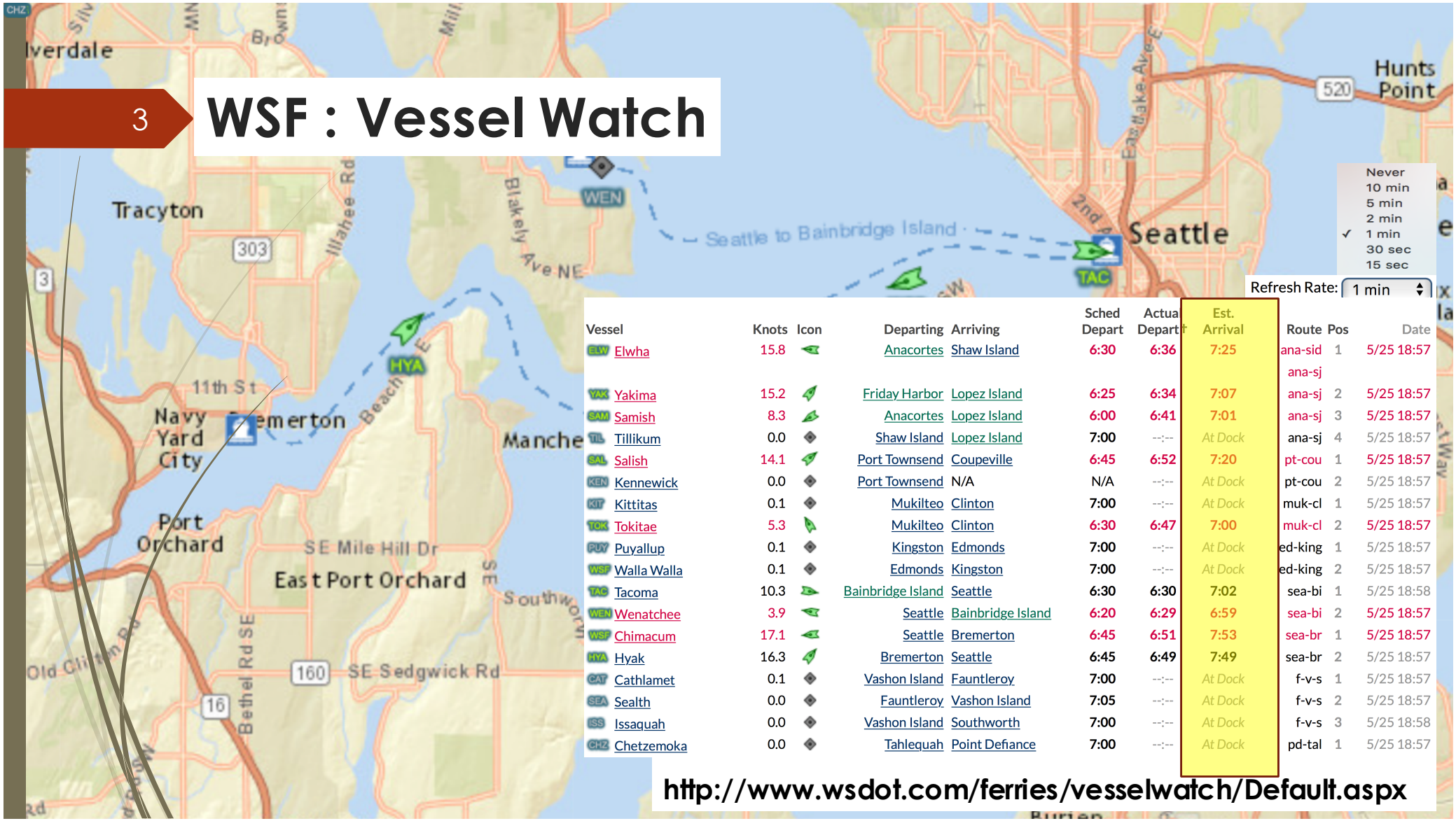
WSF : Washington State Ferry

- Nation's largest ferry system
  - Serving eight counties
  - International :Vancouver Island, British Colombia, Canada
- 10.5 million vehicles (2016)
- 24.2 million riders (2016)
- 22 auto-passenger ferries
- 20 terminals (dock stations)



Source :http://www.wsdot.wa.gov/NR/rdonlyres/6C78A08B-19A1-4919-B6E6-E9EF83E6376D/116404/WSFFactSheet2017_FINAL1.pdf

3 WSF : Vessel Watch



Vessel	Knots	Icon	Departing	Arriving	Sched Depart	Actua Depart	Est. Arrival	Route	Pos	Date
ELW Elwha	15.8		Anacortes	Shaw Island	6:30	6:36	7:25	ana-sid	1	5/25 18:57
								ana-sj		
YAK Yakima	15.2		Friday Harbor	Lopez Island	6:25	6:34	7:07	ana-sj	2	5/25 18:57
SAM Samish	8.3		Anacortes	Lopez Island	6:00	6:41	7:01	ana-sj	3	5/25 18:57
TIL Tillikum	0.0		Shaw Island	Lopez Island	7:00	--:--	At Dock	ana-sj	4	5/25 18:57
SAL Salish	14.1		Port Townsend	Coupeville	6:45	6:52	7:20	pt-cou	1	5/25 18:57
KEN Kennewick	0.0		Port Townsend	N/A	N/A	--:--	At Dock	pt-cou	2	5/25 18:57
KIT Kittitas	0.1		Mukilteo	Clinton	7:00	--:--	At Dock	muk-cl	1	5/25 18:57
TOK Tokitae	5.3		Mukilteo	Clinton	6:30	6:47	7:00	muk-cl	2	5/25 18:57
PUY Puyallup	0.1		Kingston	Edmonds	7:00	--:--	At Dock	ed-king	1	5/25 18:57
WSF Walla Walla	0.1		Edmonds	Kingston	7:00	--:--	At Dock	ed-king	2	5/25 18:57
TAC Tacoma	10.3		Bainbridge Island	Seattle	6:30	6:30	7:02	sea-bi	1	5/25 18:58
WEN Wenatchee	3.9		Seattle	Bainbridge Island	6:20	6:29	6:59	sea-bi	2	5/25 18:57
WSF Chimacum	17.1		Seattle	Bremerton	6:45	6:51	7:53	sea-br	1	5/25 18:57
HYA Hyak	16.3		Bremerton	Seattle	6:45	6:49	7:49	sea-br	2	5/25 18:57
CAT Cathlamet	0.1		Vashon Island	Fauntleroy	7:00	--:--	At Dock	f-v-s	1	5/25 18:57
SEA Sealth	0.0		Fauntleroy	Vashon Island	7:05	--:--	At Dock	f-v-s	2	5/25 18:57
ISS Issaquah	0.0		Vashon Island	Southworth	7:00	--:--	At Dock	f-v-s	3	5/25 18:58
CHZ Chetzemoka	0.0		Tahlequah	Point Defiance	7:00	--:--	At Dock	pd-tal	1	5/25 18:57

http://www.wsdot.com/ferries/vesselwatch/Default.aspx

WSF : JavaScript Object Notation (JSON) File

www.wsdot.com/ferries/vesselwatch/Vessels.ashx 200% Search

Most Visited Getting Started WSDOT - Ferries -... New Tab


```
{"feedname" : "Ferry VesselWatch", "appurl" : "http://www.wsdot.wa.gov/ferries/vesselwatch",
"legenditemlist" : [{"string" : "", "icon" : "../images/VesselWatch/boats/mobile/lightgreen1_s_90.png",
"description" : "Vessel Underway"},{"string" : "", "icon" : "../images/VesselWatch/boats/mobile
/stopped.png", "description" : "Vessel stopped or below 0.5 knots"},{"string" : "&dagger;", "icon" : "",
"description" : "Time a vessel pulls away from dock"}], "timestamp" : "5/26/2017 6:24:22 PM",
"vessellist" :
[
{"id" : 0, "inservice" : "True", "system" : "WSF", "mmsi" : 366773060, "vesselID" : 8, "name" : "Elwha",
"route" : "ANA-SID", "lastdock" : "Anacortes", "lastdock_id" : "1", "lastdock_abbrev" : "ANA", "aterm" :
"Fetching...", "aterm_id" : "", "aterm_abbrev" : "...", "eta" : "Calculating", "etaAMPM" : "", "leftdock"
: "", "departDelayed" : "", "leftdockAMPM" : "", "etaBasis" : "", "nextdep" : "6:30", "nextdepAMPM" :
"PM", "lastlat" : 0, "lastlon" : 0, "lat" : 48.507382, "lon" : -122.676895, "speed" : 0.5, "head" : 206,
"headtxt" : "Stopped", "datetime" : "5/26   18:24", "old" : 0, "icon" : "../images/VesselWatch/boats
/main/trimmed/Stopped.png", "h" : 13, "w" : 13, "xOffSet" : 0, "yOffSet" : 0, "pos" : "1", "label" :
{"icon" : "../images/VesselWatch/labels/trimmed/366773060-Stopped.png", "h" : 13, "w" : 23, "xOffSet" :
0, "yOffSet" : -13, "place" : 1}},
{"id" : 1, "inservice" : "True", "system" : "WSF", "mmsi" : 366773060, "vesselID" : 8, "name" : "Elwha",
"route" : "ANA-SJ", "lastdock" : "Anacortes", "lastdock_id" : "1", "lastdock_abbrev" : "ANA", "aterm" :
"Fetching...", "aterm_id" : "", "aterm_abbrev" : "...", "eta" : "Calculating", "etaAMPM" : "", "leftdock"
: "", "departDelayed" : "", "leftdockAMPM" : "", "etaBasis" : "", "nextdep" : "6:30", "nextdepAMPM" :
"PM", "lastlat" : 0, "lastlon" : 0, "lat" : 48.507382, "lon" : -122.676895, "speed" : 0.5, "head" : 206,
"headtxt" : "Stopped", "datetime" : "5/26   18:24", "old" : 0, "icon" : "../images/VesselWatch/boats
/main/trimmed/Stopped.png", "h" : 13, "w" : 13, "xOffSet" : 0, "yOffSet" : 0, "pos" : "1", "label" :
{"icon" : "../images/VesselWatch/labels/trimmed/366773060-Stopped.png", "h" : 13, "w" : 23, "xOffSet" :
0, "yOffSet" : -13, "place" : 2}},
```


http://www.wsdot.com/ferries/vesselwatch/Vessels.ashx

GOAL : Answer my questions

Questions:

- How accurate is WSF own ETA estimate?
- Can we do better than WSF own ETA estimate ?
  - What Machine Learning method is suitable for data?
  - Which Regression Model gives better accuracy?

SOLUTION :

- Retrieve JSON files and build the dataset
- Preprocess data & create features
- Select Machine Learning Regression methods
- Train model and predict ferry time until arrival (`LeftTime`)
- Evaluation of findings
  - Compare prediction with ferry real `LeftTime` to arrival
  - Compare prediction with WSF own ETA estimate
  - Identify reasons for ETA miscalculations if any

Related Work :

The application of machine learning techniques to the trajectory systems is a popular subject. Some examples:

- **Perera at all**: Uses ML technique Kalman Filter
  - Predict vessel trajectories not ETA
- **Parolas at all**: Uses ML techniques Support Vector Machine and Neural Networks
  - Predict ETA (by taking weather effect into consideration)
- **P.V. Shanbhag**: Uses WSF data (2016)
  - Focus in ferry pattern and visualization

Data : Source

Washington State Ferries

Build Dataset

- JSON files retrieved from WSF web site in one minute interval
- Between Jan 1, 2017 - Mar 17, 2017 (**102K JSON files**)
- All files consolidated into a single file

Raw Dataset :

- 38 variables
- 2.5 GB total volume
- 2,441,553 observations

NumPy

pandas

Data : JSON File

www.wsdot.com/ferries/vesselwatch/Vessels.ashx

Most Visited | Getting Started | WSDOT - Ferries -... | New Tab


```
{"feedname" : "Ferry VesselWatch", "appurl" : "http://www.wsdot.wa.gov/ferries/vesselwatch",
"legenditemlist" : [{"string" : "", "icon" : "../images/VesselWatch/boats/mobile/lightgreen1_s_90.png",
"description" : "Vessel Underway"},{"string" : "", "icon" : "../images/VesselWatch/boats/mobile
/stopped.png", "description" : "Vessel stopped or below 0.5 knots"},{"string" : "&dagger;", "icon" : "",
"description" : "Time a vessel pulls away from dock"}], "timestamp" : "5/26/2017 6:24:22 PM",
"vessellist" :
[
{"id" : 0, "inservice" : "True", "system" : "WSF", "mmsi" : 366773060, "vesselID" : 8, "name" : "Elwha",
"route" : "ANA-SID", "lastdock" : "Anacortes", "lastdock_id" : "1", "lastdock_abbrev" : "ANA", "aterm" :
"Fetching...", "aterm_id" : "", "aterm_abbrev" : "...", "eta" : "Calculating", "etaAMPM" : "", "leftdock"
: "", "departDelayed" : "", "leftdockAMPM" : "", "etaBasis" : "", "nextdep" : "6:30", "nextdepAMPM" :
"PM", "lastlat" : 0, "lastlon" : 0, "lat" : 48.507382, "lon" : -122.676895, "speed" : 0.5, "head" : 206,
"headtxt" : "Stopped", "datetime" : "5/26   18:24", "old" : 0, "icon" : "../images/VesselWatch/boats
/main/trimmed/Stopped.png", "h" : 13, "w" : 13, "xOffSet" : 0, "yOffSet" : 0, "pos" : "1", "label" :
{"icon" : "../images/VesselWatch/labels/trimmed/366773060-Stopped.png", "h" : 13, "w" : 23, "xOffSet" :
0, "yOffSet" : -13, "place" : 1}},
{"id" : 1, "inservice" : "True", "system" : "WSF", "mmsi" : 366773060, "vesselID" : 8, "name" : "Elwha",
"route" : "ANA-SJ", "lastdock" : "Anacortes", "lastdock_id" : "1", "lastdock_abbrev" : "ANA", "aterm" :
"Fetching...", "aterm_id" : "", "aterm_abbrev" : "...", "eta" : "Calculating", "etaAMPM" : "", "leftdock"
: "", "departDelayed" : "", "leftdockAMPM" : "", "etaBasis" : "", "nextdep" : "6:30", "nextdepAMPM" :
"PM", "lastlat" : 0, "lastlon" : 0, "lat" : 48.507382, "lon" : -122.676895, "speed" : 0.5, "head" : 206,
"headtxt" : "Stopped", "datetime" : "5/26   18:24", "old" : 0, "icon" : "../images/VesselWatch/boats
/main/trimmed/Stopped.png", "h" : 13, "w" : 13, "xOffSet" : 0, "yOffSet" : 0, "pos" : "1", "label" :
{"icon" : "../images/VesselWatch/labels/trimmed/366773060-Stopped.png", "h" : 13, "w" : 23, "xOffSet" :
0, "yOffSet" : -13, "place" : 2}},
```


http://www.wsdot.com/ferries/vesselwatch/Vessels.ashx

Data : Variable

```
"id"                : 1,
"inservice"         : "True",
"system"            : "WSF",
"vesselID"          : 2,
"name"              : "Chelan",
"lastdock"          : "Friday Harbor",
"lastdock_id"       : "10",
"lastdock_abbrev": "FRH",
"aterm"             : "Lopez Island",
"aterm_id"          : "13",
"aterm_abbrev"      : "LOP",
"datetime"          : "3/16   06:15",
"leftdock"          : "5:57",
"leftdockAMPM"      : "AM",
"departDelayed"     : "N",
"timestamp": "3/16/2017 6:15:57 AM"
```

```
"lat"     : 48.555708,
"lon"     : -122.918723,
"speed"   : 16.7,
"head"    : 41,
"headtxt": "NE",
"old"     : 0,
"mmsi"    : 366709770,
"route"   : "ANA-SJ",
"eta"     : "6:27",
"etaAMPM": "AM",
"etaBasis": "Vessel Chelan
   departed Friday Harbor going
   to Lopez and using vessel
   Kitsap closest location data
   from Mar 15 2017  4:11PM"
```

http://www.wsdot.com/ferries/vesselwatch/Vessels.ashx

Data : Cleaning

- Remove problem data

```
inservice    == True
Headtxt      != "Stopped"
aterm_abrev != ""
```

- Date-time format for all time variable (Y-m-d h:m:s)
- Grouping Data by Trip and Route
- Omit wrong Trips and missing Trips

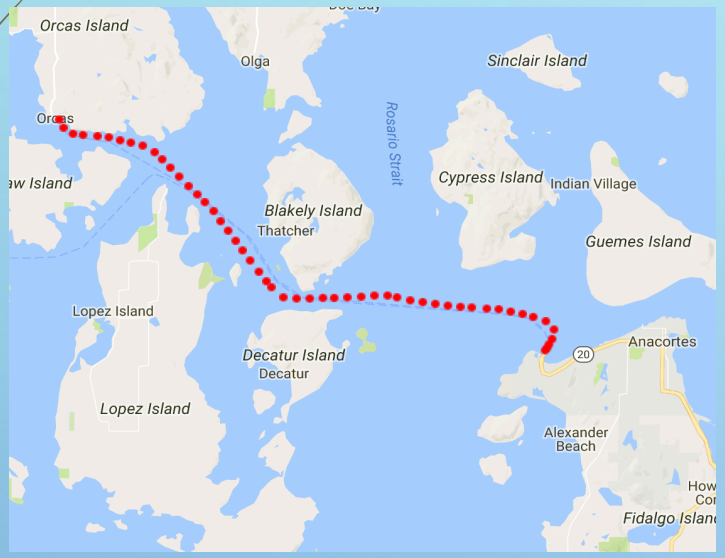
685K volume and 667,061 observations

(approx. **73 %** records omitted)

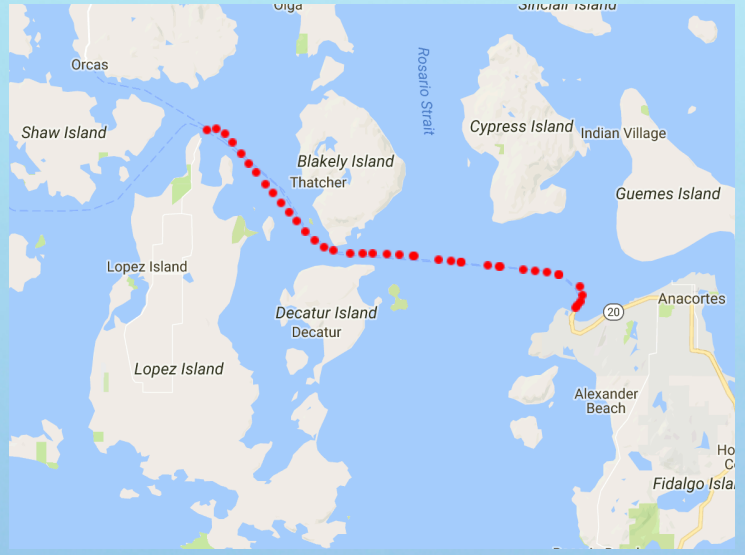
Data : Cleaning – Wrong Trips Omitted

Trips have different route than on the record, omitted

On Record: ORI - LOP



On Record: LOP - SHI

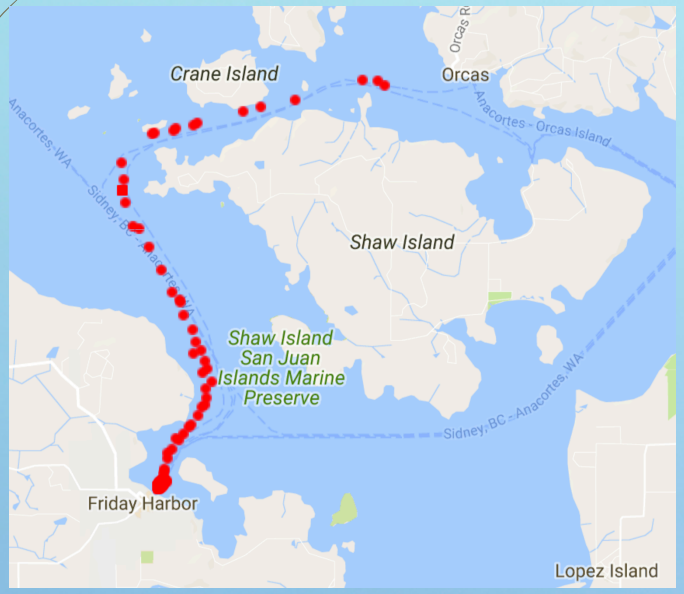


Data : Cleaning – Missing Trip Omitted

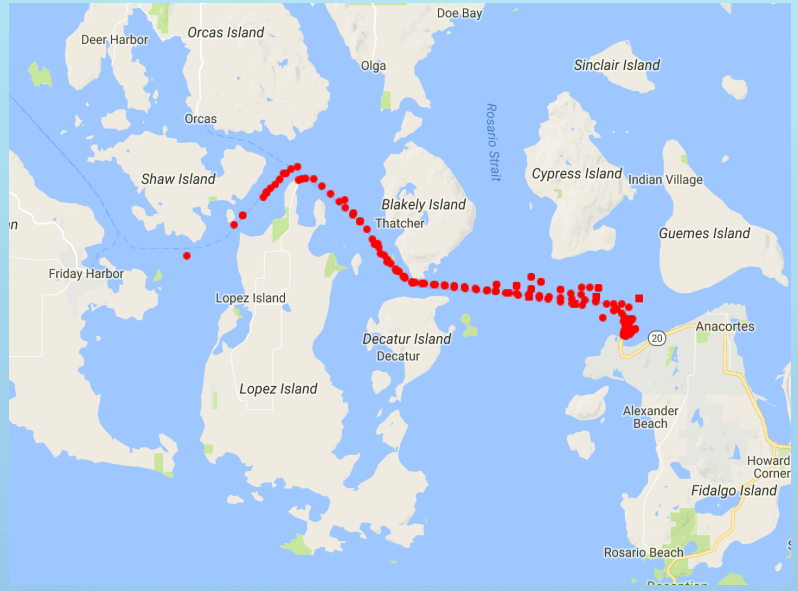
Trip Arrival time calculated by trip last record

- Missing Trips omitted

FRH-ORI



ANA - FRH



Methodology : Generating New Labels

- `SpeedAvg` Route speed average
- `n_SpeedTripAvg` Trip speed average
- `n_SpeedCumAvg` Trip speed cumulative average
- `RunStdDev` Trip running standard deviation
- `TimeStampInSec` The time of the day by seconds
- `DistLastdock` Haversine distance to last dock
- `DistAterm` Haversine distance to arrival dock aterm
- `n_dist_past` Trip cumulative distance the ferry past
- `Arrival Time` Calculated by Trip last record `datetime + 1`
  - `ArrivalTime` : Time of the ferry arrives
  - `Datetime` : Time of the record generated

Methodology : Model Development

- Input Variables
  - `DistAterm` Distance from Current location to arrival dock `aterm`
  - `Speed` Ferry's current speed
  - `TimePast` Trip time past in minutes since started
  - `RunStdDev` Trip running standard deviation
  - `TimeStampInSec` The time of the day by seconds

- Prediction :
  - `LeftTime = ArrivalTime – CurrentTime`

Methodology : Regressions & error calculation

Regression

- Linear Regression
- K-Nearest Neighbors Regression
- Random Forest Regression

Error Algorithm used for Evaluation ETA prediction with real data

- Score (R^2 coefficient of determination) : A measure of how well future samples are likely to be predicted by the model
- Correlation coefficients
- Mean Absolute Error
- Root Mean Squared Error

NumPy

matplotlib

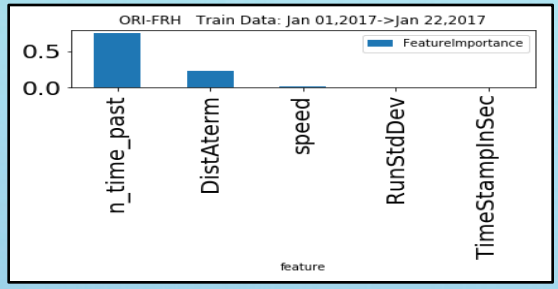
pandas

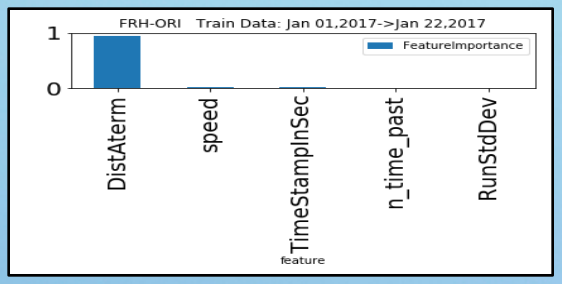
scikit learn

Findings : Python-Sklearn Tool `feature_importance_`

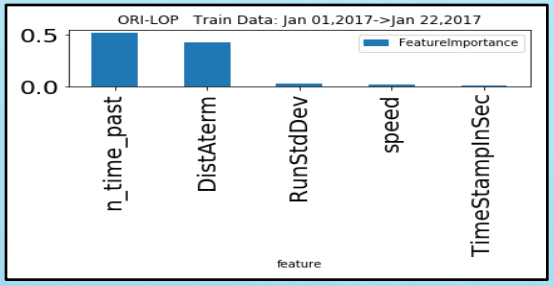
- Shows importance of features
- Uses the entire training data to calculate the informative variable

Same route - reverse

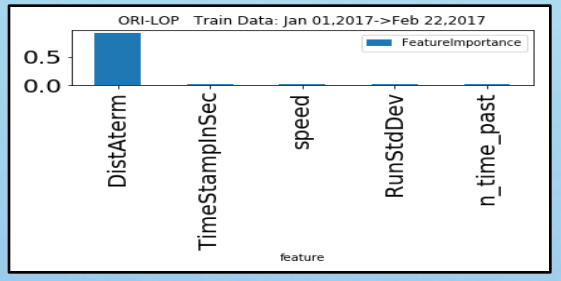




Training data different date



Jan 1 – Jan 22



Jan 1 – Feb 22

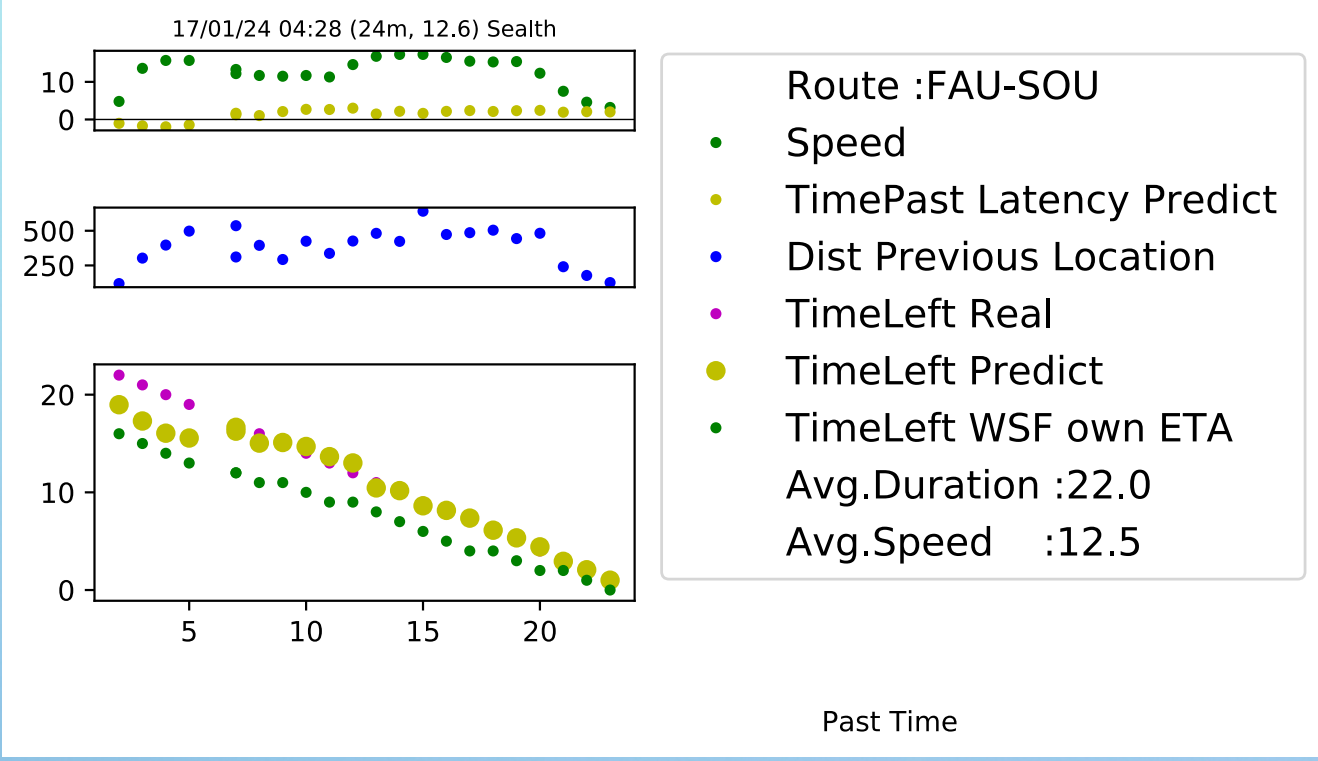
Findings : Trip Base Comparison

Predicted `LeftTime`, WSF own ETA and Real `LeftTime`

Speed / Latency

Distance from Previous Location

Real Predict WSF own ETA



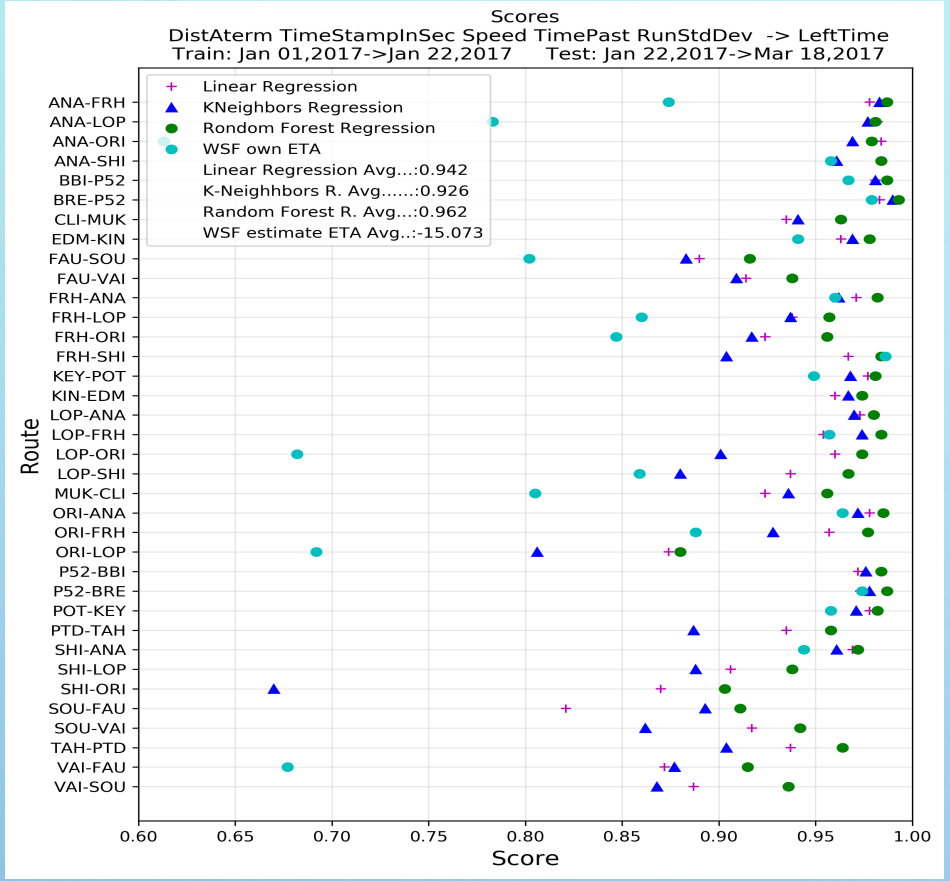
Findings : Miscalculated WSF own ETA estimate

Last Dock	Aterm	Total Record	Duration Average	eta Max
ANA	FRH	4	67	393
ANA	LOP	9	41	348
ANA	ORI	4	52	370
CLI	MUK	796	15	695
FRH	ORI	94	39	213
KIN	EDM	61	24	352
LOP	ANA	29	44	786
PTD	TAH	187	13	375
SHI	ORI	93	10	475
VAI	SOU	263	13	540

datetime	2/13 07:44
inservice	True
name	Chelan
lastdock_abbrev	ANA
aterm_abbrev	ORI
leftdock	**7:41**
leftdockAMPM	**AM**
eta	**1:51**
etaAMPM	**PM**
etaTripDur	**370**
TripAvg	52
lat	48.5165
lon	-122.68
etaBasis	Vessel Chelan departed Anacortes going to Orcas and using vessel **Chelan** closest location data from **Feb 12 2017 10:05AM**

datetime	**2/12 10:05**
inservice	**True**
name	Chelan
lastdock_abbrev	ANA
aterm_abbrev	**...**
leftdock	**NaN**
leftdockAMPM	**NaN**
eta	Calculating
etaAMPM	**NaN**
headtxt	Stopped
lat	48.5075
lon	-122.677
etaBasis	**NaN**

Findings : Regression R^2 Scores



Findings : Regression R^2 Score & Process Time

	Linear Regression		K-Neighbors Regression		Random Forest Regression		Random Forest Regression `Feature_Importances_`	
			n_neighbors=5, weights='uniform'		n_estimators=500 (The number of trees in the forest) max_features=None (The number of feature to consider) oob_score=True (whether to use out-of-bag samples) n_jobs= -1 (The number of jobs to run in parallel)			
Route	**Score**	**Train Predict Duration**	**Score**	**Train Predict Duration**	**Score**	**Train Predict Duration**	**Score**	**Train Predict Duration**
ANA-FRH	**0.98**	0.001"	**0.98**	0.44"	**0.99**	2.47"	**0.98**	4.28"
FAU-SOU	**0.89**	0.001"	**0.88**	0.16"	**0.92**	1.24"	**0.91**	2.80"
FAU-VAI	**0.91**	0.003"	**0.91**	0.91"	**0.94**	4.78"	**0.94**	9.12"

Score : A measure of how well future samples are likely to be predicted by the model

Findings : Random Forest Score and ETA

	Random Forest with Sklearn Feature Importances						Random Forest					ETA			
	Input Features: Feature_Importance_						Input Features: DistAterm, TimeStampInSec Speed, TimePast, SpeedRunDev								
	Predict : LeftTime						Predict : LeftTime								
Route	Score	Corr. Coeff.	Mean Absolute error	Root mean squared error	Train Predict Dur.	Feature_Importance_ SelectFromModel(clf, threshold=0.01)	Score	Corr. Coeff.	Mean absolute error	Root mean squared error	Train Predict Duration	Score	Correlation coefficient	Mean absolute error	Root mean squared error
ANA-FRH	**0.984**	0.99	1.5	2.39	4.28"	DistAterm,n_time_past	**0.987**	0.99	1.21	2.15	2.47"	**0.87**	0.95	1.83	6.71
BBI-P52	**0.977**	0.99	0.94	1.35	11.38"	DistAterm	**0.987**	0.99	0.68	1.03	7.69"	**0.97**	0.99	1.06	1.63
BRE-P52	**0.982**	0.99	1.42	2.09	12.90"	DistAterm	**0.993**	1.00	0.88	1.31	8.60"	**0.98**	0.99	1.58	2.23
CLI-MUK	**0.959**	0.98	0.53	0.80	8.23"	DistAterm,speed,RunStdDev	**0.963**	0.98	0.50	0.76	4.68"	**(131.99)**	0.19	5.18	45.89

Discussion, Future Work :

- Errors on WSF own ETA calculation usually happens for short routes. Results in incorrect arrival estimate.
- We don't know exactly how WSF calculate ETA estimates, but from `etaBasis`, we know which records they are using for this calculation. In our project we are already omitting those problem records.
- Future Work : Weather factors, ferry traffic and number of riders and vehicles can be taken into consideration.

Conclusion :

- Random Forest is the most accurate of the three algorithms used and has better prediction than WSF's own ETA estimate
- Washington State Ferry has to correct their system to select right data for ETA estimate reference

Thank You!