



ΔΗΜΟΚΡΙΤΟΣ
ΕΘΝΙΚΟ ΚΕΝΤΡΟ ΕΡΕΥΝΑΣ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ



ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
UNIVERSITY OF PIRAEUS

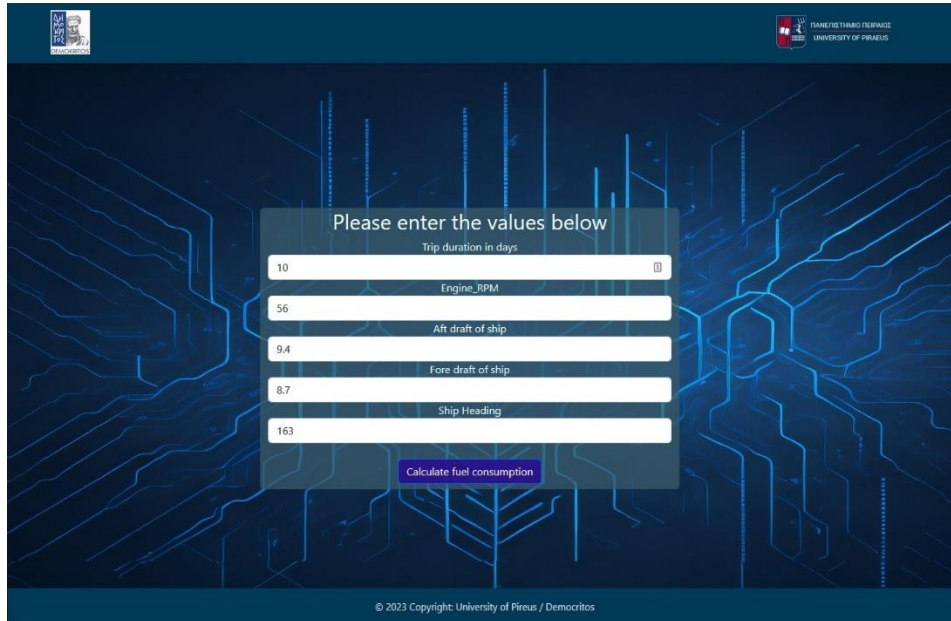


Report εργασίας ML

Αλεξίου Κυριάκος ΑΜ: 2303

3 Φεβρουαρίου, 2024

Η παρούσα εργασία αποτελεί την υλοποίηση μίας web εφαρμογής με την χρήση της οποίας μπορεί να γίνει μία πρόβλεψη της ημερήσιας, καθώς και της συνολικής κατανάλωσης καυσίμου ενός εμπορικού πλοίου κατά την πλεύση του με συγκεκριμένη κατάσταση φόρτωσης και σταθερή πορεία και ταχύτητα. Ένα μοντέλο μηχανικής μάθησης έχει εκπαιδευτεί με χρήση ιστορικών στοιχείων του υπό προσομοίωση πλοίου και χρησιμοποιείται στο back end μιας της εφαρμογής για να κάνει την πρόβλεψη κατανάλωσης. Ο χρήστης εισάγει ως δεδομένα την διάρκεια (σε μέρες) του σκέλους του ταξιδιού, τις επιθυμητές στροφές μηχανής, πρωραίο και πρυμναίο βύθισμα (που αντιστοιχούν στην κατάσταση φόρτωσης) και την κατεύθυνση του πλοίου και του επιστρέφεται η τιμή της συνολικής κατανάλωσης καυσίμου καθώς και η ημερήσια κατανάλωση σε μορφή διαγράμματος. Στα σχ.1 και σχ.2 φαίνεται παράδειγμα της αρχικής σελίδας εισαγωγής δεδομένων και της σελίδας πρόβλεψης αντίστοιχα.



Σχ.1 Σελίδα εισαγωγής δεδομένων



Σχ.2 Αποτελέσματα πρόβλεψης

Η εφαρμογή δημιουργήθηκε σε περιβάλλον python με χρήση του Django framework. Όλα τα dependencies καθώς και βοηθητικά προγράμματα που θα αναλυθούν παρακάτω, βρίσκονται αναρτημένα στο παρακάτω repo:

https://github.com/kiriakos2004/ml_assign_Democritos.git

Περιγραφή πλοίου / δεδομένα εκπαίδευσης

Τα δεδομένα του πλοίου που χρησιμοποιήθηκαν για την εκπαίδευση του μοντέλου παραχωρήθηκαν από την εταιρεία Laskaridis Shipping Co. Ltd. και αφορούν ένα BULK CARRIER εκτοπίσματος 80000DWT. Τα βασικά χαρακτηριστικά του πλοίου φαίνονται στον πίνακα1.

ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ ΠΛΟΙΟΥ	
Εκτόπισμα	80000 DWT
Συνολικό μήκος	229 m
Μέγιστο πλάτος	32,26 m
Χωρητικότητα Κ.Μηχανής	4069.44 Lt
Αριθμός χρόνων	2
Αριθμός κυλίνδρων	6
Διάμετρος προπέλας	6,95 m
Αριθμός πτερυγίων προπέλας	5

Πιν.1 Βασικά χαρακτηριστικά πλοίου

Τα δεδομένα εκπαίδευσης αφορούν ένα σύνολο 46 ημερών (από 01/04/2022 έως 16/5/2022) και λήφθηκαν με συχνότητα δειγματοληψίας 0,016Hz (ένα instance ανά λεπτό). Στο χρονικό αυτό διάστημα συλλέχθηκαν συνολικά 65868 δείγματα από 63 διαφορετικές παραμέτρους (attributes). Τα ανωτέρω δεδομένα συλλέγονται από το αυτόματο σύστημα επιτήρησης της εγκατάστασης πρόωσης του πλοίου, από ναυτιλιακά βοηθήματα καθώς και από weather providers με τους οποίους η πλοιοκτήτρια εταιρεία έχει σύμβαση. Στον πίνακα2 φαίνονται τα διαφορετικά attributes του training dataset.

DATETIME	MAIN ENGINE SCAVENGE AIR RECEIVER TEMPERATURE	AIR PRESSURE AT SEA LEVEL
AFT DRAFT	TURBOCHARGER LUB OIL INLET PRESSURE	AIR TEMPERATURE AT 10M
FORE DRAFT	TURBOCHARGER LUB OIL INLET TEMPERATURE	EASTWARD CURRENT
HEADING	THRUST BEARING TEMPERATURE	WAVE HEIGHT
INTERMEDIATE SHAFT BEARING TEMP	MIDDLE DRAFT P	WAVE DIRECTION
LIST	MIDDLE DRAFT S	WAVE LENGTH
LONGITUDINAL GROUND SPEED	RATE OF TURN	NORTHWARD CURRENT
LONGITUDINAL WATER SPEED	SPEED OVER GROUND	SEA TEMPERATURE
MAIN ENGINE SCAVENGE AIR PRESSURE	SPEED THROUGH WATER	SIGNIFICANT WAVE HEIGHT
PROPELLER SHAFT REVOLUTIONS	STARBOARD RUDDER SENSOR	SWELL WAVE HEIGHT
MAGNETIC COURSE OVER GROUND	STERN TUBE BEARING TEMPERATURE	SWELL WAVE DIRECTION
MAGNETIC VARIATION	THRUST MAIN BEARING TEMP	SWELL WAVE LENGTH
ME AXIAL VIBRATION	TRANSVERSE GROUND SPEED	SWELL SIGNIFICANT WAVE HEIGHT
ME FUEL CONSUMPTION	TRIM	WIND RELATIVE DIRECTION
FUEL OIL INLET PRESSURE	COURSE OVER GROUND	WIND RELATIVE SPEED
FUEL OIL INLET TEMPERATURE	WATER DEPTH	WINDWAVE WAVE HEIGHT
MAIN ENGINE FUEL INDEX	WIND ANGLE	WINDWAVE WAVE DIRECTION
ME RPM	WIND SPEED	WINDWAVE WAVE LENGTH

Πιν.2 Training dataset attributes

Από αυτά τα δεδομένα το “ME FUEL CONSUMPTION” ορίστηκε ως label εφόσον αντιστοιχεί στην κατανάλωση καυσίμου της κύριας μηχανής.

Προ επεξεργασία δεδομένων εκπαίδευσης

Στο πλαίσιο του περιορισμού του υπολογιστικού κόστους εκπαίδευσης καθώς και της επαύξησης της δυνατότητας πρόβλεψης του τελικού μοντέλου εφαρμόστηκαν κάποιες μέθοδοι προ επεξεργασίας του dataset. Αρχικά και με βάση το domain expertise, που τυγχάνει να κατέχεται από τον δημιουργό της εφαρμογής, απαλείφθηκαν τα attributes “DATETIME, MAGNETIC COURSE OVER GROUND, MAGNETIC VARIATION, MAIN ENGINE FUEL INDEX, MAIN ENGINE SCAVENGE AIR RECEIVER TEMPERATURE, TURBOCHARGER LUB OIL INLET PRESSURE, TURBOCHARGER LUB OIL INLET TEMPERATURE” καθώς δεν συντελούν στον καθορισμό της κατανάλωσης της Κ.Μηχανής και αποτελούν θόρυβο.

Λόγω της χαμηλής σχετικά ταχύτητας κίνησης των εμπορικών πλοίων αυτού του τύπου δεν παρουσιάζονται γρήγορες και μεγάλες μεταβολές μεγεθών. Σύμφωνα και με τις επιταγές του ISO-15016 του 2015, προς αποφυγή spikes που μπορεί να προκληθούν από ελαττωματικούς αισθητήρες εκτελέστηκε ομαλοποίηση των δεδομένων με χρήση moving average και χρονικό “παράθυρο” 10 λεπτών.

Αναλύοντας τις διακυμάνσεις των λοιπών attributes μεταξύ τους απαλείφθηκαν και τα attributes “PROPELLER SHAFT REVOLUTIONS, LONGITUDINAL GROUND SPEED, LONGITUDINAL WATER SPEED, TRIM” καθώς παρουσιάζουν ιδιαίτερα υψηλό correlation με έτερα attributes που διατηρούνται στο dataset και επομένως κρίθηκε πως δεν συντελούν στην διεύρυνση της “γνώσης” του μοντέλου. (Η συνάρτηση vis_attr() του αρχείου ml_training.py χρησιμοποιήθηκε για αυτό το σκοπό).

Παρόλο που αλγόριθμοι μηχανικής μάθησης όπως ο svm regressor δεν επηρεάζονται ιδιαίτερα από τις διακυμάνσεις της τάξης μεγέθους μεταξύ των attributes, τα δεδομένα του train dataset κανονικοποιήθηκαν με χρήση της συνάρτησης StandardScaler() της βιβλιοθήκης scikit-learn της python (το fit εκτελέστηκε μόνο στα training δεδομένα προς αποφυγή πληροφορίας στο test dataset). Πειράματα έγιναν με και χωρίς χρήση της StandardScaler.

Το train dataset περιείχε ελλείπουσες τιμές. Τα instances που είχαν ελλείπουσες τιμές στο label απορρίφθηκαν ενώ για τις περιπτώσεις που η ελλείπουσα τιμή βρισκόταν σε κάποιο από τα attributes του train dataset χρησιμοποιήθηκε η στρατηγική data imputation με χρήση της μέσης τιμής του συγκεκριμένου attribute.

Τέλος, έγινε χρήση της συνάρτησης mutual_info_regression() της βιβλιοθήκης scikit-learn με σκοπό τον επιπλέον περιορισμό του πλήθους των attributes του train dataset. Μετά από πειραματισμούς επιλέχθηκε να διατηρηθεί ένας αριθμός από τα 34 διαφορετικά attributes (Το αρχείο “Mutual_Information_Figure.jpg” παρουσιάζει τα αποτελέσματα εφαρμογής της συνάρτησης στο dataset). Πειράματα έγιναν με και χωρίς χρήση της mutual_info_regression.

Αλγόριθμοι / διαδικασία εκμάθησης / τεστ

Με σκοπό την δημιουργία του καλύτερου δυνατού μοντέλου επιλέχθηκαν να δοκιμαστούν οι επιδόσεις ενός εκπροσώπου από σχεδόν κάθε μεγάλη κατηγορία regression αλγορίθμων. Αρχικά και με σκοπό τον καθορισμό ενός threshold επιδόσεων, με βάση το οποίο θα αξιολογηθούν οι λοιποί αλγόριθμοι, χρησιμοποιήθηκε ένας linear regressor. Οι αλγόριθμοι που δοκιμάστηκαν πλέον του ανωτέρω είναι ο SVM regressor, ο GradientBoosting regressor καθώς και ο Bagging regressor (με χρήση kNN estimator) ώστε να καλυφθούν και τεχνικές

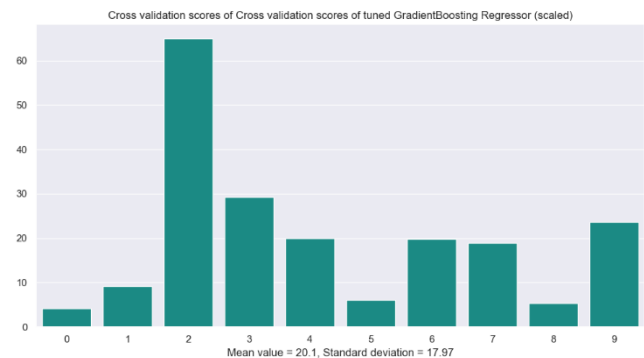
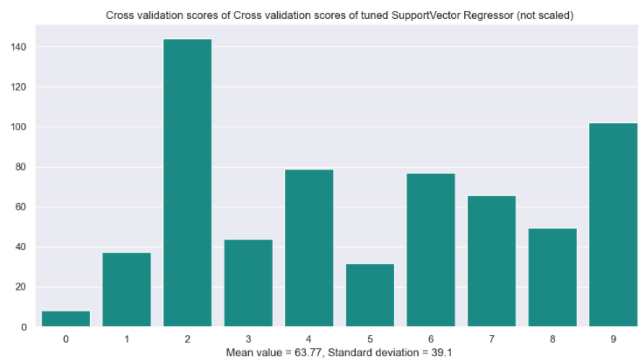
boosting / bagging. Το αρχείο ml_training.py χρησιμοποιήθηκε για την εκπαίδευση του τελικού μοντέλου και την αποθήκευση του.

Ως στρατηγική hyperparameter tuning επιλέχθηκε το Grid search σε ένα εύρος τιμών των σημαντικότερων hyperparameters κάθε ενός από τους αλγορίθμους. Οι συναρτήσεις svc_regressor() και gb_regressor() εκτελούν grid search στις hyperparameters που φαίνονται στον πίνακα 3.

SVC Regressor		GradientBoosting Regressor	
hyperparameter	values	hyperparameter	values
C	0.1, 1, 10, 100, 200, 300, 500 , 1000	Learning rate	0.01, 0.1
kernel	'rbf', 'linear', 'poly'	Max depth	5 , 10, 20, 30
-	-	L2_regularization	0.1, 0.01, 0.001

Πιν.3 Εύρος των Hyperparameters

Οι ανωτέρω συναρτήσεις εκτελούν **cross-validated grid-search** στις hyperparameters του ανωτέρω πίνακα 3 με χρήση 10 folds. Οι βέλτιστες hyperparameters, οι οποίες είναι bold στον πίνακα, επιλέχθηκαν όχι μόνο με βάση την καλύτερη τιμή της μετρικής (RMSE) αλλά και με βάση την “σταθερότητα” (μικρή διακύμανση) των αποτελεσμάτων σε αυτά τα 10 folds. Όταν επιλέχθηκαν οι καλύτερες hyperparameters με βάση την ανωτέρω διαδικασία, συγκρίθηκαν τα αποτελέσματα των tuned αλγορίθμων ξανά με βάση την απόλυτη τιμή της μετρικής καθώς και την συνεκτικότητα των αποτελεσμάτων. Τα αποτελέσματα της διαδικασίας αξιολόγησης επιδόσεων των αλγορίθμων παρουσιάζονται στο αρχείο “results.txt” και οι γραφικές παραστάσεις τους μπορούν να αξιολογηθούν με χρήση του αρχείου “result_scores.py”. Ενδεικτικά στις παρακάτω δύο εικόνες παρουσιάζονται οι καλύτερες επιδόσεις των tuned SVC Regressor και GradientBoosting Regressor.



Από τα παραπάνω διαγράμματα φαίνεται η υπεροχή του **GradientBoosting Regressor** και για αυτό και επιλέχθηκε εκείνος. Επίσης ως συμπέρασμα μπορούμε να εξάγουμε πως στον GradientBoosting Regressor είχαμε βελτίωση επιδόσεων (αν και μικρή) με χρήση normalization χωρίς ιδιαίτερη αύξηση του υπολογιστικού κόστους ενώ ο SVC Regressor δεν επηρεάστηκε θετικά (πράγμα αναμενόμενο). Τέλος η χρήση της συνάρτησης Mutual Information δεν επέφερε κάποια αξιοσημείωτη βελτίωση των επιδόσεων αν και βελτίωσε ελαφρώς το training time.

Το τελικό μοντέλο (tuned GradientBoosting Regressor) δοκιμάστηκε ως προς την ικανότητα γενίκευσης με την χρήση ξεχωριστού dataset το οποίο αφορούσε σε δεδομένα διαφορετικού ταξιδιού του πλοίου που έγινε σε μεταγενέστερο χρόνο υπό λίγο διαφορετικές καιρικές συνθήκες. Δόθηκε ιδιαίτερη προσοχή ώστε να υπάρχει σχεδόν όλη η διαθέσιμη γκάμα

στροφών μηχανής στο test dataset ώστε να αποκτηθεί μια όσο το δυνατόν πληρέστερη εικόνα των δυνατοτήτων προσέγγισης της κατανάλωσης καυσίμου από το τελικό μοντέλο. Η γκάμα στροφών του test dataset είναι μεταξύ 3.15 και 83.6 rpm που ουσιαστικά αποτελούν το σύνολο των διαθέσιμων στροφών μηχανής. Οι επιδόσεις στο test dataset ήταν οι εξής:

Maximum error: 169.59 Lt/min

Mean absolute error: 18.87 Lt/min *

*Αξίζει να σημειωθεί πως μία τέτοια μηχανή έχει μια κατανάλωση της τάξης των 150.000 με 200.000 λίτρων καυσίμου την ημέρα.

Υπόλοιπα δεδομένα εισαγωγής

Τα δεδομένα εισαγωγής στο μοντέλο μπορούν να χωριστούν σε τρεις μεγάλες κατηγορίες:

- ❖ Τα δεδομένα τα οποία “χαρακτηρίζουν” το ταξίδι όπως είναι πχ οι στροφές της μηχανής. Τα δεδομένα αυτά είναι εκείνα τα οποία επιλέχθηκαν να εισάγονται από τον χρήστη στη εφαρμογή καθόσον αποτελούν επιλογές που εκείνος θα αξιολογήσει με την βοήθεια της εφαρμογής. Τα δεδομένα που καλείται να εισάγει ο χρήστης είναι :
 - **Trip Duration** (εκτιμώμενη διάρκεια ταξιδιού)
 - **Engine RPM** (στροφές μηχανής, σημειώνεται πως τα πλοία αυτού του είδους διατηρούν τις στροφές σταθερές καθ’ όλη τη διάρκεια του ταξιδιού)
 - **Aft Draft**
 - **Fore Draft**
 - **Heading** (σημειώνεται πως τα πλοία αυτού του είδους διατηρούν συνήθως σταθερή την πορεία τους στο μεγαλύτερο μέρος του ταξιδιού)
- ❖ Τα δεδομένα που αποτελούν λειτουργικές παραμέτρους της εγκατάστασης του πλοίου που σε πραγματικές συνθήκες παρέχονται από το σύστημα παραμετρικής παρακολούθησης πλατφόρμας του πλοίου. Η εφαρμογή παράγει simulated δεδομένα αυτής της κατηγορίας με χρήση των συναρτήσεων “create_var_inherited” και “create_var_dummy” του αρχείου “var_creation.py”. Και οι δύο ανωτέρω συναρτήσεις κάνουν χρήση του φυσικού μοντέλου του πλοίου. Η πρώτη συνάρτηση δημιουργεί δεδομένα με χρήση των inputs που εισαγάγει ο χρήστης ενώ η δεύτερη επιλέγει μέσα από μία κατανομή λειτουργικών χαρακτηριστικών του πλοίου.
- ❖ Τα δεδομένα που συνήθως παρέχονται στο πλοίο από εξωτερικούς φορείς όπως π.χ. δεδομένα καιρού. Η εφαρμογή παράγει τυχαία simulated δεδομένα αυτής της κατηγορίας με χρήση της συνάρτησης “create_random_env_vars” επίσης του αρχείου “var_creation.py”. (Τα δεδομένα αυτά θα μπορούσαν να εισέρχονται στο μοντέλο με χρήση ενός API μιας weather εφαρμογής).