ε.δε.μ²
ΕΠΙΣΤΗΜΗ ΔΕΔΟΜΕΝΩΝ & ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΕΡΓΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

Επεξηγησιμότητα Τεχνητής Νοημοσύνης - eXplainable AI (XAI)

Ορισμοί, Intrinsic & Model-Agnostic XAI Methods

PI (Permutation Feature Importance)

SHAP (Shapley Additive exPlanations)

LIME (Local Interpretable Model Agnostic Explanation)

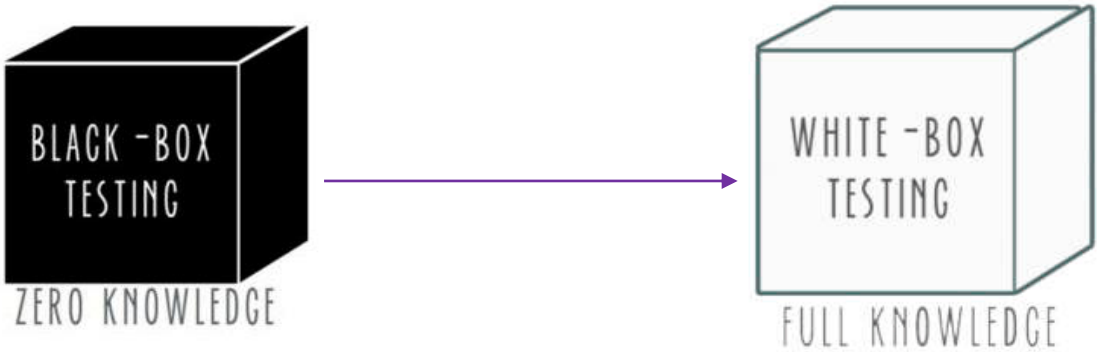
καθ. Βασίλης Μάγκλαρης
maglaris@netmode.ntua.gr
www.netmode.ntua.gr

Αίθουσα 002, Νέα Κτίρια ΣΗΜΜΥ
Τρίτη 6/6/2023

ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

Αναγκαιότητα Μεθόδων eXplainable Artificial Intellignece (XAI) (1/2)

https://link.springer.com/book/10.1007/978-3-030-68640-6



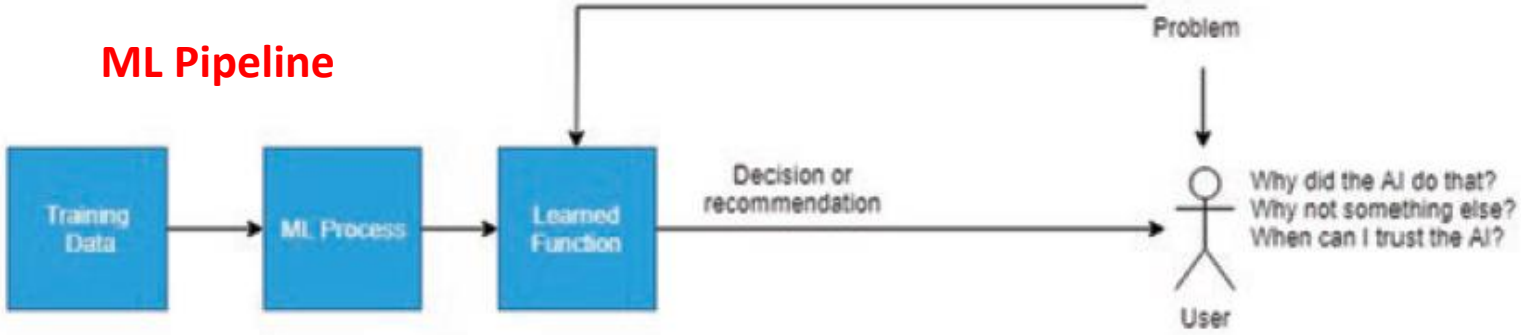
- Η ακρίβεια (***accuracy***) δεν είναι αρκετή για την επιλογή μοντέλου ***Machine Learning*** (***ML***)
- Δυσπιστία χρηστών - σχεδιαστών - αναλυτών - ρυθμιστών πολιτικής σε μη ερμηνεύσιμα (***uninterpretable***) συστήματα αποφάσεων τύπου ***black-box***
- Αναγκαιότητα μεθόδων ***eXplainable Artificial Intelligence*** (***XAI***) ως προς τα κριτήρια λήψης αποφάσεων & ρύθμισης μοντέλου προς σχεδιαστές - χρήστες συστημάτων ***ML***
- Δικαιολόγηση αποφάσεων συστημάτων ***ML*** σε ερωτήσεις χρηστών - πελατών για προσωπικά ζητήματα που τους αφορούν (***local interpretations***)
- Ανάγκη για ***συγκριτική αξιολόγηση features*** και δικαιολόγηση σχεδιαστικών ***επιλογών*** μοντέλου, παραμέτρων/υπερπαραμέτρων κλπ.
- Ανάπτυξη εργαλείων αποτύπωσης σημασίας (***importance***) και συσχετίσεων (***correlations***) χαρακτηριστικών (***features***) σε γραφικό περιβάλλον, φιλικό προς τους χρήστες

ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

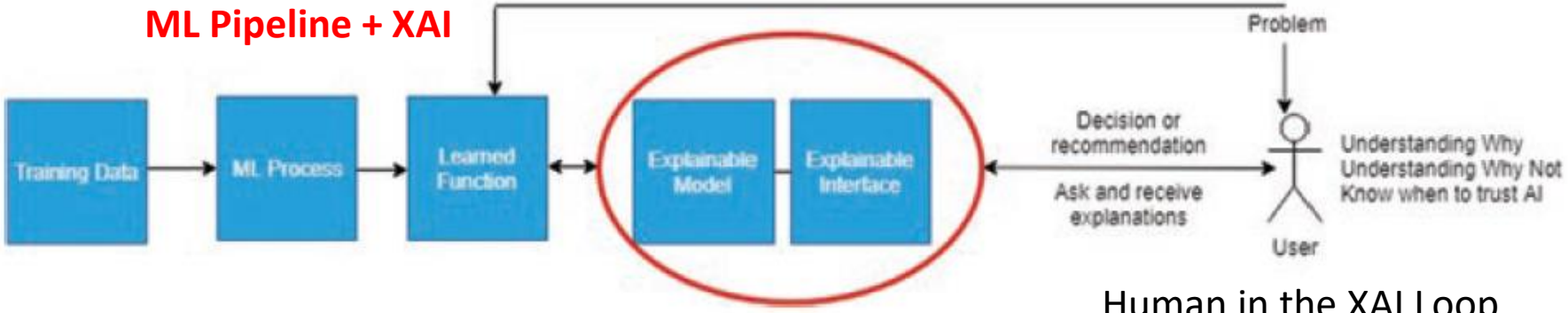
Αναγκαιότητα Μεθόδων eXplainable Artificial Intellignece (XAI) (2/2)

https://link.springer.com/book/10.1007/978-3-030-68640-6

ML Pipeline



ML Pipeline + XAI



Human in the XAI Loop

XAI

ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

Ορισμοί Ερμηνευσιμότητας - Interpretability & Επεξηγησιμότητας - Explainability

https://link.springer.com/book/10.1007/978-3-030-68640-6

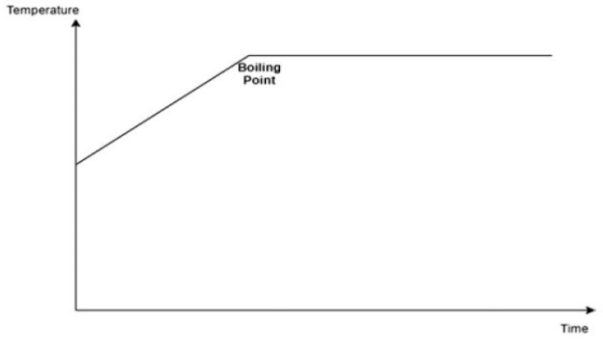
- **Interpretability:** Possibility of understanding the mechanics of an ML model, but not necessarily knowing why
- **Explainability:** Understanding why

Question	Interpretability	Explainability
Which are the most important features that are adopted to generate the prediction or classification?	✓	✓
How much the output depends on small changes in the input?	✓	✓
Is the model relying on a good range of data to select the most important features?	✓	✓
What are the criteria adopted to come across the decision?	✓	✓
How would the output change if we put different values in a feature not present in the data?	✗	✓
What would happen to the output if some feature or data had not occurred?	✗	✓

Explainability →

Interpretability
(but not the opposite)

- Model describing the process of boiling water
- Task: Predict temperature given time



Temperature can be predicted, but the physics of the boiling point cannot be directly explained

Παράδειγμα Διαφοράς:
Μοντέλο Βρασμού Νερού

- Interpretable: YES
- Explainable: NO

ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

Ταξινόμηση Τεχνικών eXplainable AI (Taxonomy of XAI Techniques)

https://link.springer.com/book/10.1007/978-3-030-68640-6

- **Intrinsic vs. Post-Hoc Models**
  - **Intrinsic**: Εγγενώς ερμηνεύσιμα μοντέλα (π.χ. βάρη σε Linear Regression, Gini Index για διαμόρφωση Decision Trees, επίδραση features σε ταξινόμηση βάση K-Nearest Neighbors)
  - **Post-hoc**: Η επεξήγηση έπεται της μάθησης - ρύθμισης παραμέτρων και αφορά στο δείγμα test
- **Model Agnostic vs. Model Specific Post-Hoc Models**
  - **Model Agnostic**: Επεξηγήσεις ανεξάρτητα από τη δομή μοντέλου αποφάσεων (black-box) με αξιολόγηση εξόδου ρυθμισμένου συστήματος
  - **Model Specific**: Επεξηγήσεις για συγκεκριμένες παραμέτρους γνωστού μοντέλου αποφάσεων
- **Global vs. Local Explainability**
  - **Global**: Ερμηνείες (interpretations) για το σύνολο του δείγματος
  - **Local**: Ερμηνείες (interpretations) για συγκεκριμένα δειγματικά στοιχεία

ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

Ταξινόμηση Τεχνικών eXplainable AI (Taxonomy of XAI Techniques)

https://link.springer.com/book/10.1007/978-3-030-68640-6

- **Intrinsic vs. Post-hoc Models**
  - **Intrinsic**: Εγγενώς ερμηνεύσιμα μοντέλα (π.χ. βάρη σε Linear Regression, Gini Index για διαμόρφωση Decision Trees, επίδραση features σε ταξινόμηση βάση K-Nearest Neighbors)
  - **Post-hoc**: Η επεξήγηση έπεται της μάθησης - ρύθμισης παραμέτρων και αφορά στο δείγμα test
- **Model Agnostic vs. Model Specific Post-hoc Models**
  - **Model Agnostic**: Επεξηγήσεις ανεξάρτητα από τη δομή μοντέλου αποφάσεων (black-box) με αξιολόγηση εξόδου ρυθμισμένου συστήματος
  - **Model Specific**: Επεξηγήσεις για συγκεκριμένες παραμέτρους γνωστού μοντέλου αποφάσεων
- **Global vs. Local Explainability**
  - **Global**: Ερμηνείες (interpretations) για το σύνολο του δείγματος
  - **Local**: Ερμηνείες (interpretations) για συγκεκριμένα δειγματικά στοιχεία

Παράδειγμα Intrinsic XAI Model: Προβλέψεις Πωλήσεων Smartphones / Ηλικία Αγοραστών

LINEAR REGRESSION MODEL

Απλή Ερμηνεία, Προβληματική Ακρίβεια

Fig. 2.6 A linear monotonic function gives simple, ready-to-go explanations with one global characteristic, the variation of purchases for one unit of age

NON-LNEAR REGRESSION MODEL

Δύσκολη Ερμηνεία - Προσθήκη Local Μοντέλων
Βελτίωση Ακρίβειας - Κίνδυνος Overfitting

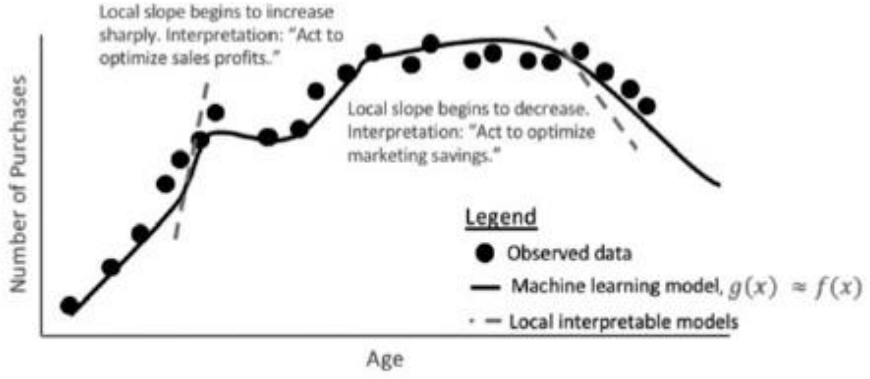


Fig. 2.7 With nonlinear non-monotonic function, we lose a global easily explainable model in favor of an accuracy improvement

ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

Model-Agnostic Μέθοδοι XAI

https://link.springer.com/book/10.1007/978-3-030-68640-6

- **PI** (Permutation Importance)
- **SHAP** (Shapley Additive exPlanations)
- **LIME** (Local Interpretable Model agnostic Explanation)

ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

Permutation Importance - PI

https://link.springer.com/book/10.1007/978-3-030-68640-6

- Επιστρέφει κατά σειρά τα πιο σημαντικά χαρακτηριστικά (***features***) του δείγματος για προβλέψεις (***predictions***) του συστήματος ***ML*** με διαδοχικές ανακατατάξεις (***Permutations***) των ***features*** δειγματικών στοιχείων εισόδου
- Μέθοδος κατάταξης των ***features*** (**Permutation Importance - PI**) για ***Post-Hoc,*** **Global Explainability** στο δείγμα ***Testing*** (όχι στο δείγμα μάθησης – ***Training Dataset***)
- Συγκρίνει την έξοδο του συστήματος με ανακατάξη (***reshuffling***) της σειράς εισόδου ενός χαρακτηριστικού (***feature***) δειγματικού στοιχείου ***testing***. Αν παρατηρείται σημαντική διαφορά το συγκεκριμένο χαρακτηριστικό θεωρείται σημαντικό
- Δεν συγκρίνει τιμές των ***features***, απλά τις ταξινομεί κατά σειρά ανάλογα με την επίδραση στην έξοδο (***prediction***) του συστήματος
- Δεν αποκαλύπτει συσχετίσεις (***correlations***) των features
- Η επιρροή ενός χαρακτηριστικού (***feature***) που κατατάχτηκε σαν σημαντικό στην αναζήτηση πρόβλεψης μπορεί να απεικονίζεται μέσω ***Partial Dependence Plot – PDP***

Η γρήγορη κατάταξη των ***features*** **ΚΑΙ** στο ***Training Dataset*** μπορεί να επιταχύνει τη μάθηση αλγορίθμων ***ML*** και να μειώσει κινδύνους ***overfitting*** με την διαγραφή μη σημαντικών χαρακτηριστικών (ή και με τιμές αρνητικές) κατά την ρύθμιση (***regularization***) των ***datasets***

ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

Shapley Additive exPlanations – SHAP (1/3)

https://link.springer.com/book/10.1007/978-3-030-68640-6

- Η μεθοδολογία **SHAP** αφορά σε εκ των υστέρων (***post-hoc***) ερμηνεία της επιρροής χαρακτηριστικών (***features***) δειγματικών στοιχείων εισόδου (***input sample points***) στην ***έξοδο*** ενός συστήματος ***ML***
- Αναφέρεται σε **συγκριμένο δειγματικό στοιχείο εισόδου** (***local prediction***) μετά τη διαδικασία μάθησης (***post-hoc***), με βάση τα δεδομένα εισόδου - εξόδου και όχι τη δομή του μοντέλου που έχει ήδη ρυθμιστεί (***model agnostic***)
- Βασίζεται στις τιμές ***Shapley*** που εισήγαγε το 1953 ο ***Lloyd S. Shapley*** https://en.wikipedia.org/wiki/Lloyd_Shapley για ανάλυση της συνεισφοράς M παικτών σε συνεργατικά στοχαστικά παίγνια (***stochastic cooperative games***)
- Αναλογία παικτών με M ***features*** στη διαδικασία αποτίμησης της οριακής συνεισφοράς τους σε ***prediction*** στην έξοδο συστήματος ***ML***, με βάση τις εξόδους $y = f_X(S)$ σε εισόδους $\mathrm{x} \in X$ για ***features*** που περιορίζονται στο υποσύνολο S. Τα $f_X(S)$ εκτιμώνται μέσω προσομοιωμένης διαδικασίας μάθησης για δειγματικό υποσύνολο X (***background set***) στο ρυθμισμένο σύστημα ***ML*** (***post-hoc***)
- Αν φ_i είναι η συνεισφορά του ***feature*** i (***Shapley Value***) η συνολική συνεισφορά όλων είναι το άθροισμα $\sum_{i=0}^{M} \varphi_i$ (όπου $\varphi_0 = constant$) το οποίο πρέπει να επιμερισθεί στα M ***features*** ανάλογα με τις ***Shapley Values*** (θεωρούμε πως οι φ_i είναι αθροιστικές και μεταβάλλονται μονοτονικά ως προς το μέγεθος της συνεισφορά τους)
- Υπολογισμός των ***Shapley Values*** για συνεισφορά του ***feature*** i σε δειγματικό σημείο του ***test dataset***:

$$\varphi_i = \sum_{S \sqsubseteq N\{i\}} \frac{(M - |S| - 1)!\,|S|!}{M!} [f_X(S \cup \{i\}) - f_X(S)]$$

Το άθροισμα υπολογίζεται για όλα τα υποσύνολα S των ***features*** που δεν περιλαμβάνουν την i, $f_X(S)$ είναι η έξοδος με είσοδο το υποσύνολο S, ενώ το $f_X(S \cup \{i\}) - f_X(S)$ είναι η συνεισφορά της ***feature*** i

Ο συντελεστής $\frac{(M-|S|-1)!|S|!}{M!}$ αντανακλά τον αριθμό συνδυασμών υποσυνόλων των ***features***

ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

Shapley Additive exPlanations – SHAP (2/3)

https://link.springer.com/book/10.1007/978-3-030-68640-6

SHAP Force Plot: Πρόβλεψη Ομάδας Καλύτερου Παίκτη Αγώνα Uruguay – Russia

- Η πρόβλεψη με σύστημα ***ML*** (***Random Forest***) αφορά στον ***post-hoc*** προσδιορισμό της πιθανότητας ο καλύτερος παίκτης να ανήκει στην ομάδα της ***Uruguay*** (***local XAI***)
- Διερευνάται η επίδραση (***impact***) στη πρόβλεψη των χαρακτηριστικών (***feature***) εισόδου π.χ. «***αριθμός των τερμάτων της Uruguay***» συγκεκριμένου δειγματικού στοιχείου (***local XAI***)
- Με ***κόκκινο*** χρώμα παρίστανται τα ***features*** με θετική συνεισφορά, π.χ. «***αριθμός των τερμάτων της Uruguay = 3***»
- Με χρώμα ***μπλε*** τα ***features*** με αρνητική συνεισφορά
- Το μέγεθος της θετικής ή αρνητικής συνεισφοράς είναι ανάλογο με το μήκος του σχετικού τμήματος στο διάγραμμα
- Η τιμή $\mathbf{0.52}$ για την έξοδο (πιθανότητα ο καλύτερος παίκτης να ανήκει στην ***Uruguay***) αντιστοιχεί με το άθροισμα των κόκκινων τμημάτων μείον το άθροισμα των μπλε. Είναι λίγο μεγαλύτερη από το μέσο όρο $\mathbf{0.50}$ της απόλυτης αοριστίας!

Fig. 4.7 SHAP diagram that shows how the features impact on the match Uruguay-Russia. A force diagram representing how much the features change the final value. For example we see that "Goal Scored = 3" has the most impact for it pushes the final value to the right with the biggest interval (Becker 2020)

ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

Shapley Additive exPlanations – SHAP (3/3)

https://link.springer.com/book/10.1007/978-3-030-68640-6

SHAP Summary Plot: Πρόβλεψη Ομάδας Καλύτερου Παίκτη Αγώνα Uruguay – Russia

SUMMARY PLOT

- Διερευνάται η συνεισφορά (***impact***) ***features*** πολλαπλών δειγματικών στοιχείων εισόδου (σε ρυθμισμένο σύστημα ***ML***) μέσω επαναληπτικών ***local*** υπολογισμών των ***SHAP Values*** (προς ***global explainability***), με μέση ***SHAP Value*** $\mathbf{0.0}$
- Πλήθος δειγματικών στοιχείων (υποσύνολο του ***test dataset***): Αριθμός κουκίδων στο διάγραμμα
- Με ***κόκκινο*** χρώμα παρίστανται τα δειγματικά στοιχεία με **μεγάλα *feature values.*** Με χρώμα ***μπλε*** τα στοιχεία με **μικρά *feature values***

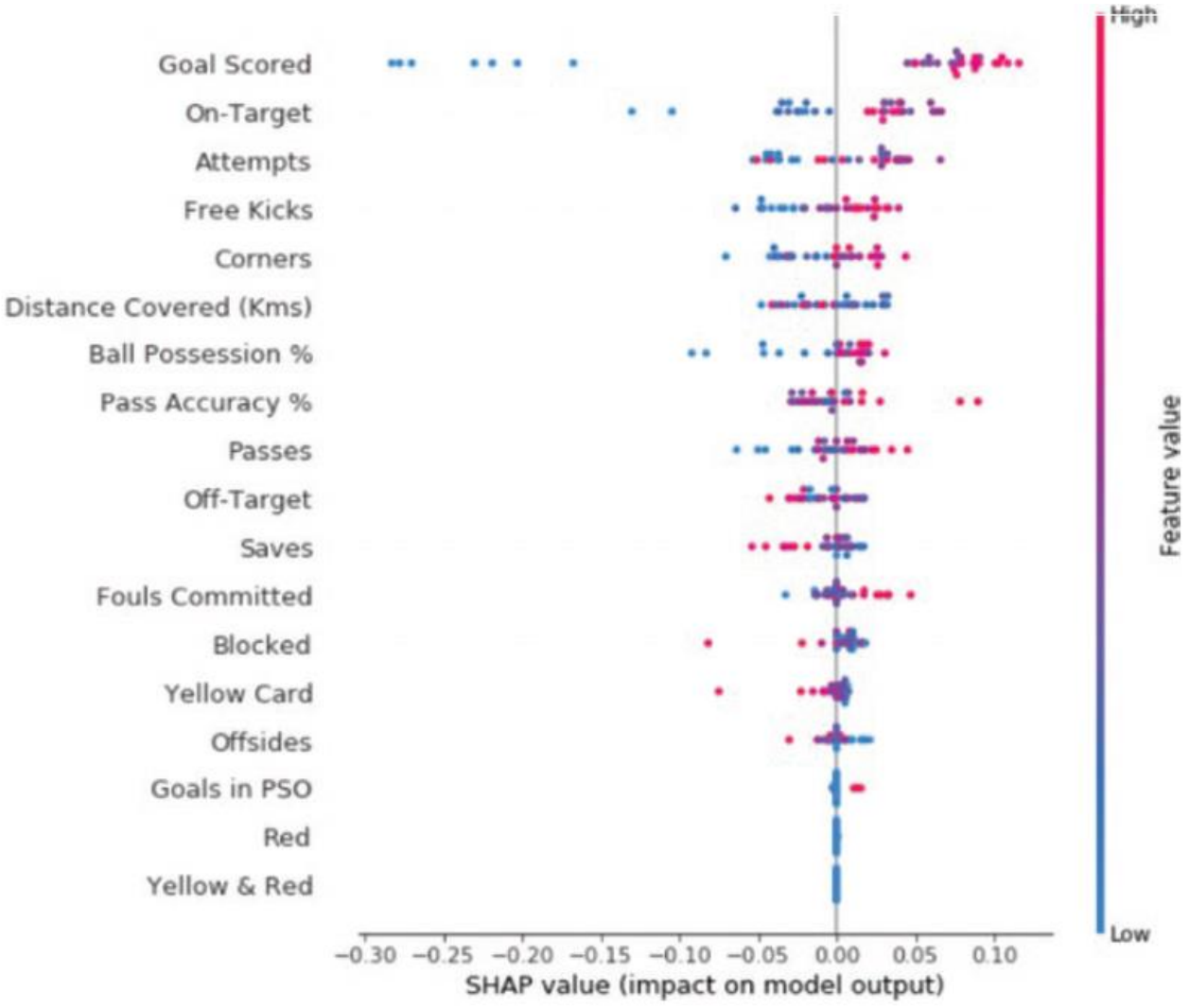


Fig. 4.8 SHAP diagram that shows the features' ranking and the related impact on the match prediction (Becker 2020)

ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

Local Interpretable Model agnostic Explanation - LIME

https://link.springer.com/book/10.1007/978-3-030-68640-6

- Τροποποίηση του δείγματος προς ***surrogate*** (αναπληρωματικά) δειγματικά στοιχεία (π.χ. με προσθήκη θορύβου ***Gauss***) για ***model agnostic*** αντιστοίχιση μοντέλων ***black-box*** με μοντέλα ***white-box*** που να διευκολύνουν την ερμηνεία ενός νέου δειγματικού στοιχείου
- Στόχος η μείωση διαστάσεων (αριθμό ***features***) λόγω απαλοιφής των μη σημαντικών χαρακτηριστικών και η υλοποίηση γραμμικών συστημάτων προβλέψεων (π.χ. ***linear regression***) που επιδέχονται ***intrinsic interpretations***
- Τα τροποποιημένα ***surrogate*** δειγματικά στοιχεία αποκτούν νέα βάρη που ευνοούν συγκεντρώσεις γύρω από το στοιχείο προς ερμηνεία (***local interpretability***), ενώ μειώνουν την επίδραση ***outliers*** για εύκολους αλγορίθμους γραμμικής ταξινόμησης
- Το τροποποιημένο δείγμα παρέχει ευκολότερη ***local interpretability*** με λιγότερα ***features*** και ικανοποιητική ακρίβεια προβλέψεων (***accuracy***) εφόσον τα κοντινά ***surrogate*** στοιχεία δεν έχουν σημαντικές αποκλίσεις (π.χ. ελάχιστο ***Mean Square Error***) από τα αρχικά

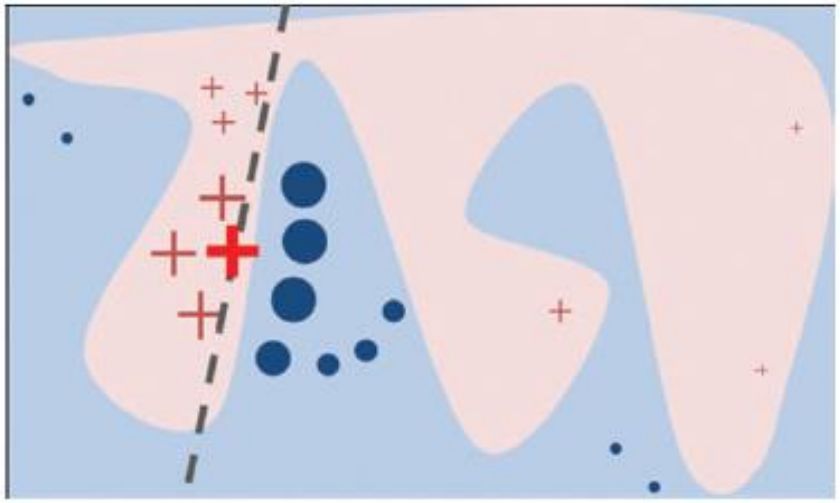


Fig. 4.9 Schematics for LIME. The class outputs of the model are circles or crosses, and the dimension reminds us of the weight, so distant points are weighted less (Ribeiro et al. 2016)

Προσεγγίζουμε το δειγματικό στοιχείο x (**Κεντρικός Κόκκινος Σταυρός**) με «κοντινά» ***surrogate*** στοιχεία, με βάρη ανάλογα με την απόσταση ώστε να προκύπτει γραμμική διαχωρισιμότητα (***linear separability***) στην περιοχή του x σε δύο κλάσεις: **Σταυροί** και **Κύκλοι**

Το ***surrogate model*** είναι ***intrinsically explainable***, τουλάχιστον στην γειτονική περιοχή του x (***local XAI***)