ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ

ΤΜΗΜΑ ΕΦΑΡΜΟΣΜΕΝΗΣ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ

1η Εργασία

Insurance Cross Sell Prediction using TF Keras

Όνομα: Δημήτρης

Επίθετο: Μανωλάκης

Α.Μ.: it1423

Περιεχόμενα

Introduction

Στόχος της παρούσας εργασίας είναι η εκπαίδευση νευρωνικών δικτυών μέσω των οποίων θα πραγματοποιηθούν προβλέψεις σχετικά με την απάντηση πελατών ασφαλιστικής εταιρίας όταν αυτή τους προσφέρει ασφάλεια αυτοκίνητου. Η αξία ενός τέτοιου μοντέλου/νευρωνικού δικτύου είναι ιδιαίτερα υψηλή για ασφαλιστικές εταιρίες οι οποίες θα έχουν την δυνατότητα να βελτιστοποιήσουν το επιχειρηματικό τους μοντέλο και την επικοινωνιακή τους στρατηγική, απευθυνόμενοι σε πελάτες που όντως ενδιαφέρονται, μεγιστοποιώντας κατά αυτόν τον τρόπο τα έσοδα τους.

Για την εκπαίδευση των νευρωνικών δικτυών θα γίνει χρήση συνόλου δεδομένων που αφορούν πελάτες ασφαλιστικής εταιρίας που έχουν ήδη απαντήσει σχετικά με το ενδιαφέρον τους για ασφάλεια αυτοκίνητου. Το σύνολο δεδομένων αποτελείται από δημογραφικά στοιχεία των πελατών (φύλο, ηλικία κ.α.), των αυτοκίνητων που διαθέτουν (ηλικία, κατάσταση κ.α.) καθώς και πληροφορίες σχετικά με το συμβόλαιο της εταιρίας με τον πελάτη (διάρκεια, τρόποι επικοινωνίας κ.α.). Τα δεδομένα αυτά έχουν εξαχθεί από το Kaggle.

Data Cleaning

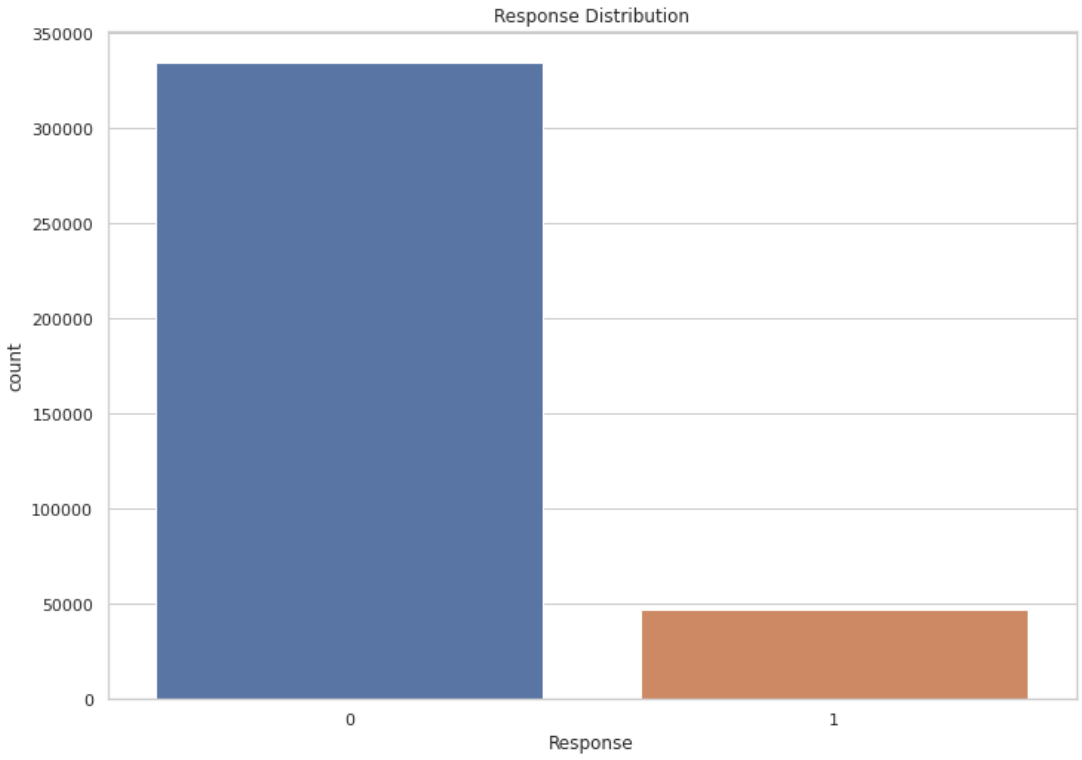
Ένα πρώτο βασικό βήμα σε μια διαδικασία ανάλυσης δεδομένων είναι ο καθαρισμός τους καθώς συχνά περιέχουν ατέλειες. Πιο συγκεκριμένα, πραγματοποιήθηκε σχετικός έλεγχος για τυχόν ελλιπής δεδομένα ή λανθασμένες τιμές στα δεδομένα, όπως για παράδειγμα ηλικίες οδηγών μικρότερες του 18.

Τα δεδομένα ήταν γενικά σε πολύ καλή κατάσταση μιας και προέρχονται από σχετικό διαγωνισμό του Kaggle. Στο βήμα αυτό αφαιρέθηκε η στήλη 'ID' (αναγνωριστικό) από κάθε εγγραφή του πίνακα των δεδομένων καθώς δεν προσφέρει κάποια πληροφορία στο μοντέλο.

Exploratory Data Analysis

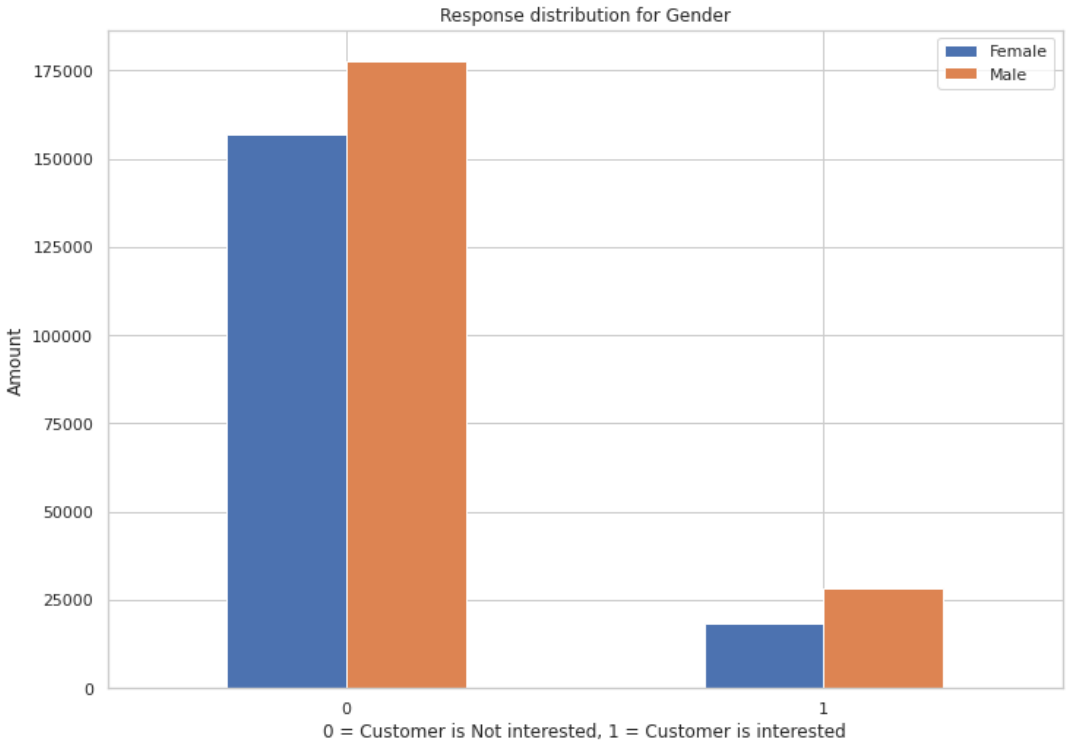
Προτού γίνει δημιουργία προβλεπτικού μοντέλου, είναι απαραίτητη η οπτικοποίηση των δεδομένων έτσι ώστε να εξαχθούν συμπεράσματα σχετικά με τη δομή των δεδομένων καθώς και για τις σχέσεις που διέπουν τα χαρακτηριστικά του dataset.

Target Value - Response



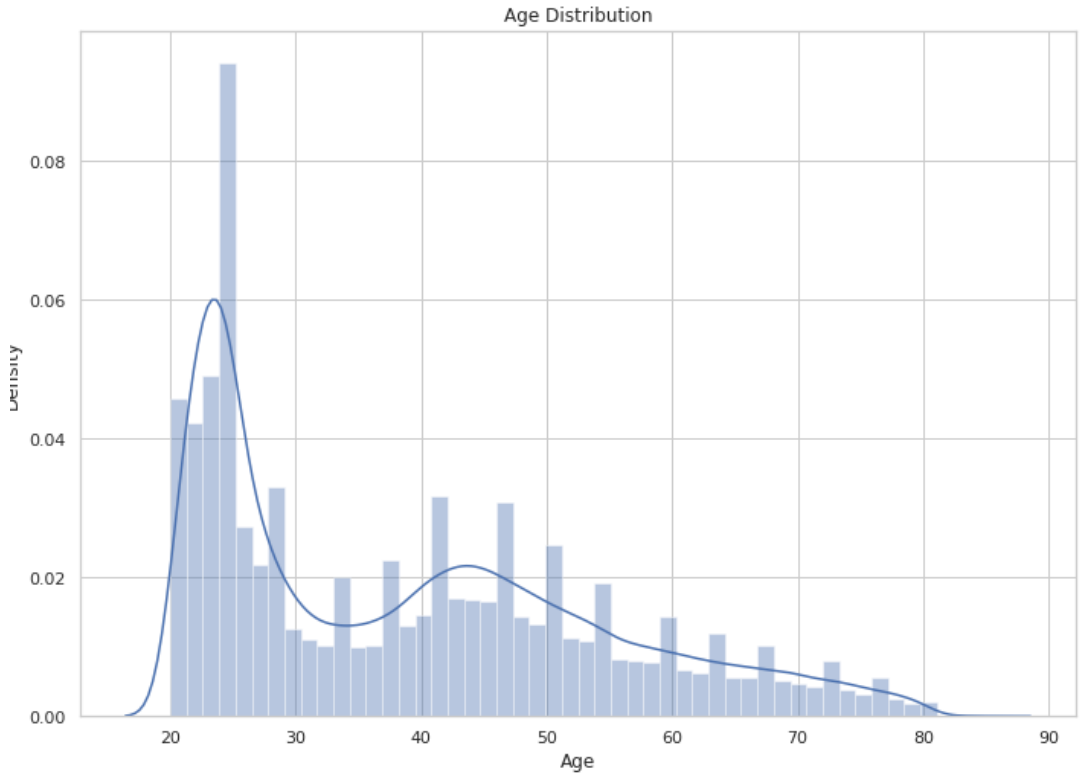
Στο παραπάνω Plot αναπαρίσταται η κατανομή των απαντήσεων των πελατών σχετικά με την επιθυμία τους να αγοράσουν ασφάλεια αυτοκίνητου (οπού 0 η απάντηση είναι αρνητική, οπού 1 η απάντηση είναι θετική). Είναι προφανές πως η πλειοψηφία των πελατών, και πιο συγκεκριμένα, το 87.74% των πελατών έχουν δώσει αρνητική απάντηση. Η έλλειψη ισορροπίας αυτή αποτελεί σημαντικό εμπόδιο στην μοντελοποίηση των δεδομένων, καθώς οι θετικές απαντήσεις είναι πολύ σπάνιες.

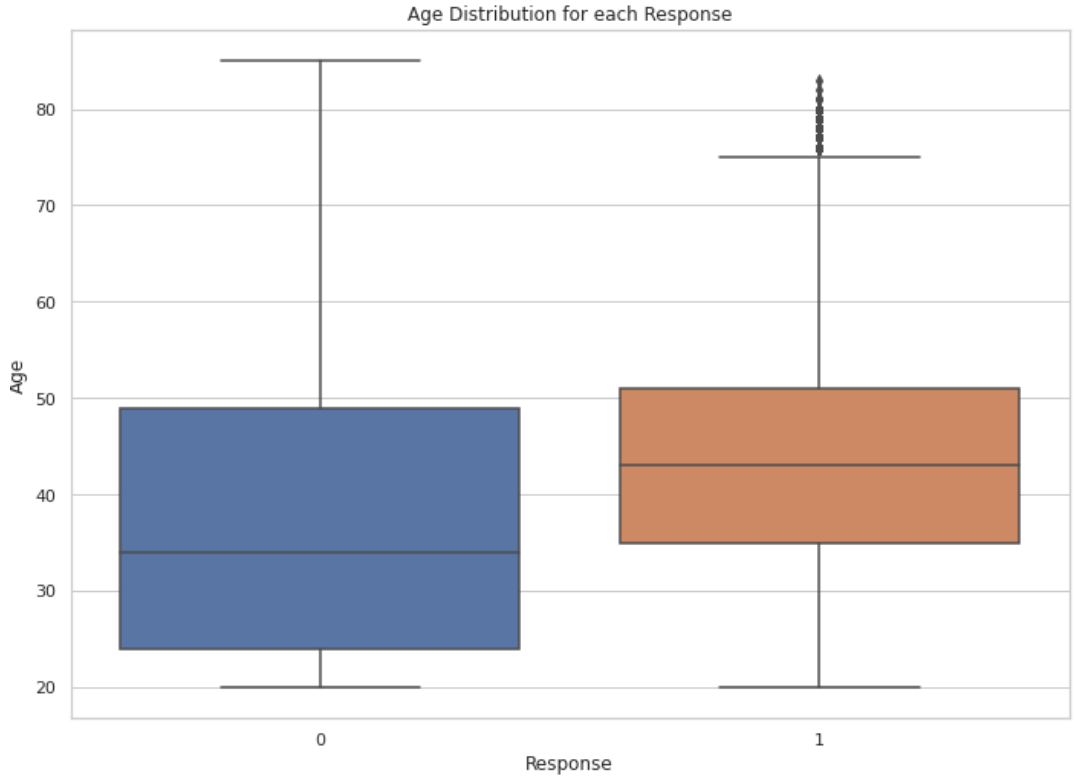
Gender



Η κατανομή των θετικών και των αρνητικών απαντήσεων αναμεσά στα δυο φυλά είναι ιδιαίτερα ισορροπημένη.

Age


Age Distribution
Density
0.08
0.06
0.04
0.02
0.00
20
30
40
50
60
70
80
90
Age

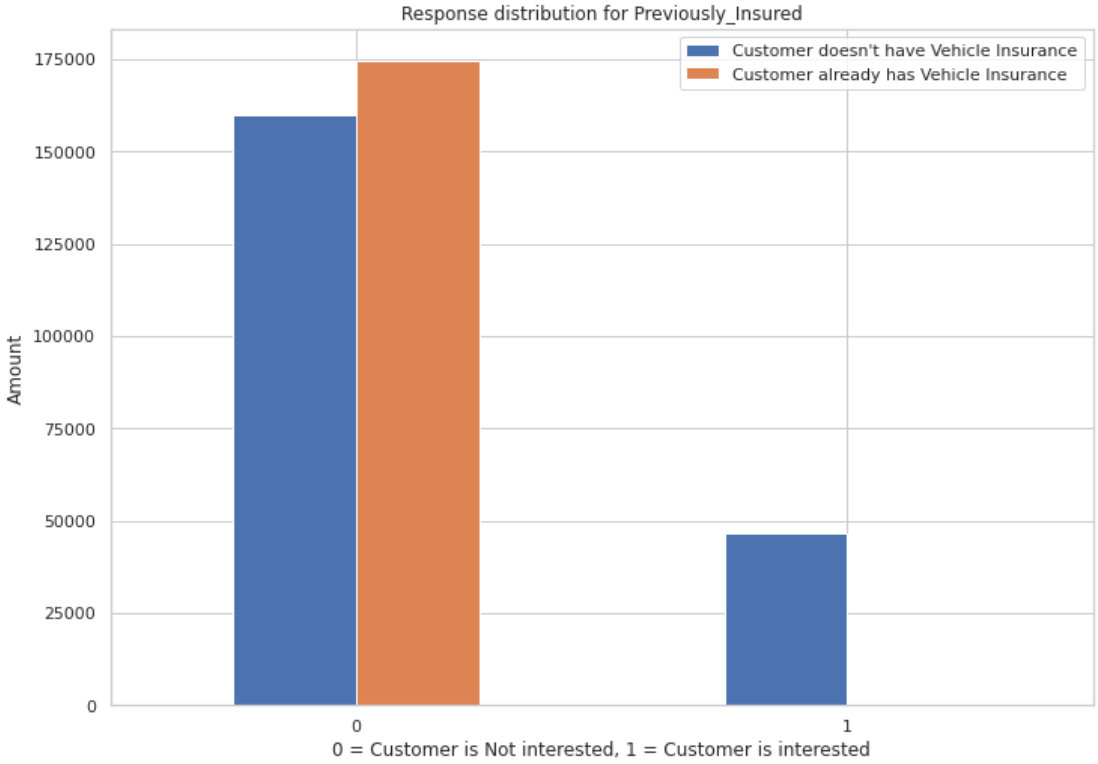

Age Distribution for each Response
Age
80
70
60
50
40
30
20
0
1
Response

Οι περισσότεροι πελάτες βρίσκονται στο ηλικιακό πλαίσιο των 20-25 ετών όπως είναι προφανές από το πρώτο Plot, ενώ στο θηκόγραμμα της δεύτερης εικόνας βλέπουμε ότι οι πελάτες που έχουν δώσει θετική απάντηση κυμαίνονται στο ηλικιακό πλαίσιο των 35-50 ετών.

Driving License

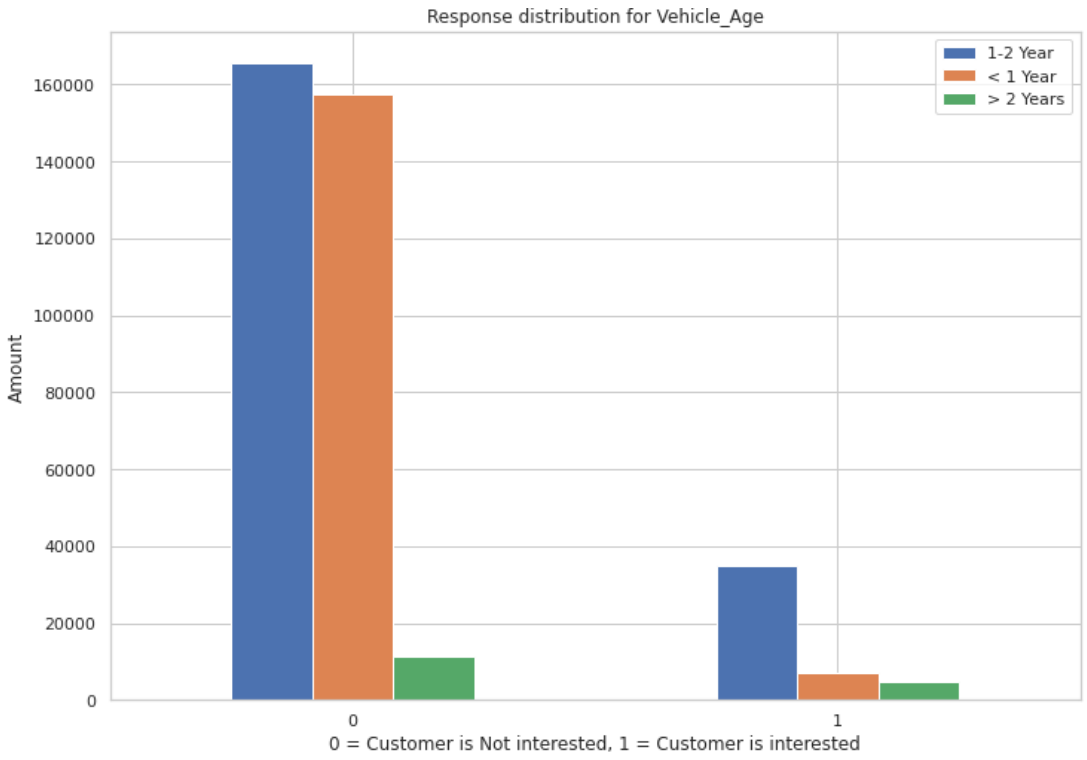
Από τους 381.000 πελάτες του dataset, μόνο 812 δεν διαθέτουν δίπλωμα οδήγησης και επομένως, για την απλοποίηση του μοντέλου το χαρακτηριστικό αυτό αφαιρέθηκε από τα δεδομένα.

Previously Insured



Οι πελάτες που δίνουν θετική απάντηση για ασφάλεια αυτοκίνητου δεν είχαν προηγουμένως κάποια άλλη ασφάλεια αυτοκίνητου.

Vehicle Age



Οι ιδιοκτήτες αυτοκίνητων ηλικίας 1-2 ετών κυριαρχούν στο dataset καθώς και στις θετικές απαντήσεις. Αξιοπερίεργο συμπέρασμα αποτελεί το γεγονός πως οι ιδιοκτήτες αυτοκίνητων ηλικίας κάτω του ενός χρόνου δείχνουν ιδιαίτερα αρνητικοί στην αγορά ασφάλειας αυτοκίνητου. Το πλήθος των οδηγών αυτοκίνητων αυτής της ηλικίας που έδωσαν θετική απάντηση είναι σχεδόν ίσο με οδηγούς αυτοκίνητων ηλικίας

μεγαλύτερης των 2 ετών, μολονότι οι τελευταίοι σπανίζουν στο σύνολο δεδομένων.

Data Preprocessing

Στο τελικό βήμα πριν την μοντελοποίηση, γίνεται η κατάλληλη προετοιμασία και προ επεξεργασία των δεδομένων ώστε να αποκτήσουν την κατάλληλη μορφή. Η προ επεξεργασία για τα περισσότερα χαρακτηριστικά του dataset ήταν τυπική:

Οι αριθμητικές μεταβλητές Region Code και Policy Sales Channel μετατράπηκαν σε ακέραιοι, ενώ οι δυαδικές μεταβλητές Gender, Previously Insured και Vehicle Damage μετατράπηκαν στην μορφή (-1,1). Η μορφή αυτή επιλέχθηκε υστέρα από σχετική έρευνα για την ιδανική αναπαράσταση δυαδικών μεταβλητών σε νευρωσικά δίκτυα οπού και εντοπίστηκε η παραπάνω, σε άρθρο του Warren Sarle (SAS Institute, Neural Networks Specialist).

Τέλος, ιδιαίτερη μεταχείριση χρειάστηκε για την μεταβλητή Vehicle Age καθώς αποτελεί ordinal categorical variable με τιμές: > 2 Years, 1-2 Years και < 1 Year. Για το encoding της μεταβλητής αυτής επιλέχθηκε η υποδιαίρεση της σε 2 νέες μεταβλητές: Vehicle Age Over Year και Vehicle Age Over 2 Years. Αυτοκίνητα ηλικίας μικρότερης τους ενός έτους οδηγούνται στο ΄διάνυσμα’:

Vehicle Age Over Year = 0 και Vehicle Age Over 2 Years = 0

Αυτοκίνητα ηλικίας στο διάστημα 1-2 έτη οδηγούνται στο ΄διάνυσμα’:

Vehicle Age Over Year = 1 και Vehicle Age Over 2 Years = 0

Αυτοκίνητα ηλικίας μεγαλύτερης των 2 ετών οδηγούνται στο ΄διάνυσμα’:

Vehicle Age Over Year = 1 και Vehicle Age Over 2 Years = 1

Η μορφή αυτή επιλέχθηκε διότι θεωρήθηκε πως αποθηκεύει την μέγιστη πληροφορία για το εκπαιδευόμενο νευρωνικό δίκτυο.

Modeling

Για την σωστή εκπαίδευση άλλα και εκτίμηση της απόδοσης του μοντέλου, το σύνολο δεδομένων διαχωρίστηκε σε 3 επιμέρους σύνολα δεδομένων: train (342.998 δείγματα), test (38.111 δείγματα) και validation set (34.300 δείγματα).

Μοντέλο #1

Προκειμένου να εξαχθεί μια πρώτη εικόνα της απόδοσης ενός νευρωνικού δικτύου στα δεδομένα

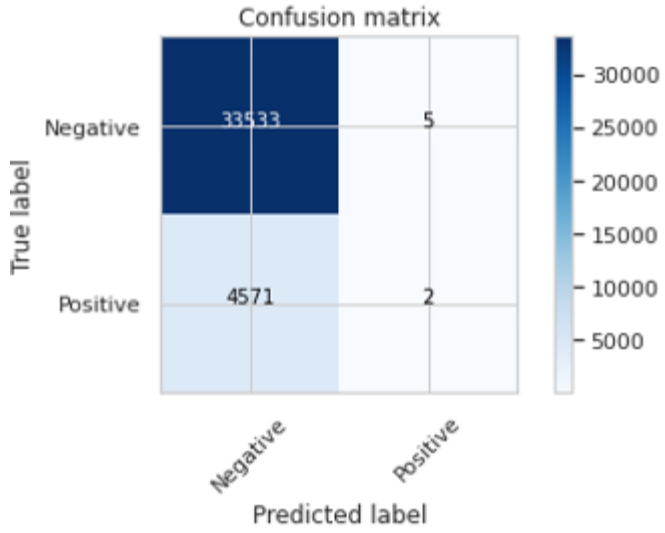
δημιουργήθηκε ένα benchmark μοντέλο το οποίο αποτελείται από δυο Hidden Layers των 16 και 32 νευρώνων οι οποίοι κάνουν χρήση της ReLU activation function. Αμέσως πριν το Output Layer έχει τοποθετηθεί Dropout Layer το οποίο αγνοεί το 50% των τυχαία επιλεγμένων νευρώνων της εισόδου του και καταπολεμά κατά αυτό το τρόπο το Overfitting. Όπως συνηθίζεται σε Binary Classifications, το Output Layer περιέχει έναν νευρώνα ο οποίος κάνει χρήση της σιγμοειδής συνάρτησης ενεργοποίησης ενώ η loss function που επιλέχθηκε για το δίκτυο είναι η Binary Crossentropy. Ως Optimizer για το δίκτυο αυτό επιλέχθηκε ο ιδιαίτερα δημοφιλής Adam optimizer με learning rate ίσο με 0.001.

Τέλος, για την επιπλέον αντιμετώπιση του Overfitting γίνεται χρήση του Early Stopping. Αν μετά από 10 εποχές δεν έχει υπάρξει βελτίωση του επιλεγόμενου metric (Validation AUC) τότε η εκπαίδευση σταματά και γίνεται restore των weights του καλυτέρου έως τότε μοντέλου. Η εκπαίδευση του μοντέλου γίνεται για 100 εποχές ή μέχρι να συμβεί Early Stopping callback, κάνοντας χρήση του training και του validation dataset.

Μετά την εκπαίδευση του, η οποία έληξε πρόωρα στις 46 εποχές, το μοντέλο πραγματοποιεί προβλέψεις στο test dataset. Οι έξοδοι του νευρωνικού δικτύου προέρχονται από την σιγμοειδής συνάρτηση ενεργοποίησης του νευρώνα στο Output Layer και επομένως αποτελούνται από τιμές αναμεσά στο 0 και

το 1. Μολονότι η σιγμοειδής συνάρτηση δεν αποτελεί συνάρτηση πυκνότητας πιθανότητας οι εξόδιο αυτοί μπορούν να ερμηνευτούν ως ένδειξη βεβαιότητας (confidence) του δικτύου για την απάντηση του εκάστοτε πελάτη (μιας και αποτελεί αθροιστική συνάρτηση πιθανότητας της λογιστικής κατανομής). Οι έξοδοι αυτές επομένως μετατρέπονται με χρήση ενός threshold της τάξης του 0.5, είτε σε 0 (Κλάση Αρνητικής Απάντησης) για εξόδους στο πεδίο [0,0.5), είτε σε 1 (Κλάση Θετικής Απάντησης) για εξόδους στο πεδίο (0.5,1]. Το threshold ελαφρώς ευνοεί την κλάση των θετικών απαντήσεων μιας και αυτές είναι σπάνιες και επομένως ακόμη και μια μικρή σχετικά βεβαιότητα του δικτύου (π.χ. 0.5) θεωρείται αρκετή.

Αν αρκεστούμε στο accuracy του μοντέλου κατά την εκπαίδευση στα validation data τότε μπορούμε να πούμε με σιγουριά πως δημιουργήθηκε ένα πολύ καλό μοντέλο καθώς πέτυχε 88% accuracy. Υστέρα από την μετατροπή των προβλέψεων στην κατάλληλη μορφή είμαστε πλέον σε θέση να περάσουμε τα αποτελέσματα του δικτύου και τις πραγματικές τιμές σε confusion matrix το οποίο θα μας δώσει μια πιο ουσιαστική εικόνα για την απόδοση του δικτύου:



Classification Report

	precision	recall	f1-score	support
Negative	0.88	1.00	0.94	33538
Positive	0.29	0.00	0.00	4573
accuracy			0.88	38111
macro avg	0.58	0.50	0.47	38111
weighted avg	0.81	0.88	0.82	38111

Το δίκτυο δυστυχώς προβλέπει μόνο τιμές της κλάσης 0 καθώς όπως προαναφέρθηκε κυριαρχεί σε μεγάλο βαθμό στα δεδομένα εκπαίδευσης. Παράλληλα όμως κυριαρχεί και στα δεδομένα ελέγχου και επικύρωσης δίνοντας αυτή την ψευδαίσθηση ενός καλού μοντέλου με accuracy ίσο με 88% ενώ η πληροφορία που κρύβεται πίσω από αυτό το νούμερο είναι ουσιαστικά πως το 88% των απαντήσεων στα δεδομένα ελέγχου είναι αρνητικές.

Το metric που καταλληλά εκφράζει το φαινόμενο αυτό είναι το Macro Average f1 score, το οποίο όπως φαίνεται στην παραπάνω εικόνα έχει τιμή ίση με 0.47.

Για τον υπολογισμό του metric αυτού συνεισφέρουν ισότιμα όλες οι κλάσεις και επομένως εντοπίζει ευκολά τέτοιου είδους ανωμαλίες στην απόδοση του μοντέλου. Στόχος της εργασίας αυτής θα είναι η μεγιστοποίηση του metric αυτού.

Τέλος, ένα άλλο metric που εκφράζει την απόδοση του μοντέλου και για τις δυο κλάσεις είναι το Area Under the Curve (AUC) και για τον λόγο αυτό, οπότε χρησιμοποιείται Early Stopping αυτό θα είναι το metric του οποίου η βελτίωση θα παρακολουθείται.

Resampling

Μια ευρέως χρησιμοποιουμένη μέθοδος για προβλήματα αυτού του είδους είναι το Resampling. Η δυνατότητα αυτή παρέχεται από τη βιβλιοθήκη της Python imblearn (στην δημιουργία της οποίας συμμετείχε ο κος Χρήστος Αριδάς, Πανεπιστήμιο Πάτρας). Η μέθοδος αυτό ισορροπεί τις κλάσεις ενός dataset αφαιρώντας δείγματα του majority class (under-sampling) ή/και πρόσθεση επιπλέον δειγμάτων στο minority class (over-sampling).

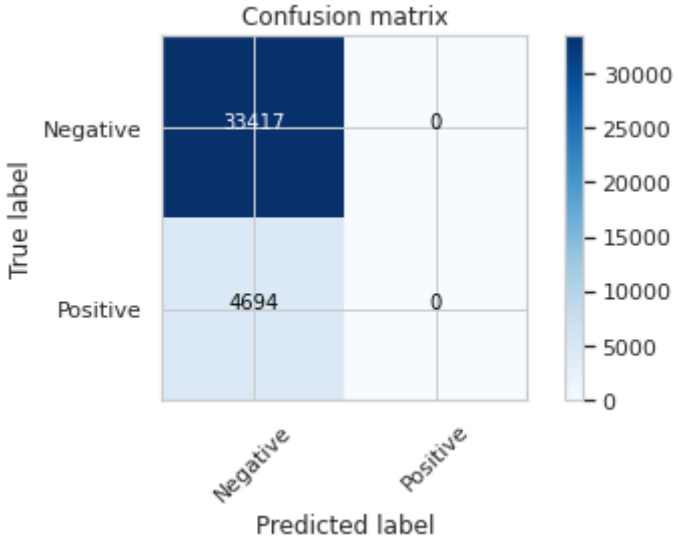
Αν και πρόκειται για μια ιδιαίτερα χρήσιμη και δυνατή μέθοδο έχει τις αδυναμίες της καθώς το over-sampling γίνεται με τυχαία αντιγραφή δειγμάτων που υπάρχουν ήδη στο Dataset και πιθανώς να οδηγήσει σε Overfitting ενώ το under-sampling γίνεται με τυχαία διαγραφή

δειγμάτων η οποία μπορεί να οδηγήσει απώλεια πληροφορίας.

Μοντέλο #2

Το μοντέλο αυτό χρησιμοποιεί την ιδιά ακριβώς αρχιτεκτονική με το μοντέλο 1 άλλα στα δεδομένα εκπαίδευσης του, και πιο συγκεκριμένα, στα δείγματα της κλάσης 0 (Αρνητικές απαντήσεις) έχει πραγματοποιηθεί under-sampling (από 270.884 δείγματα σε 200.000).

Confusion Matrix:



Classification Report

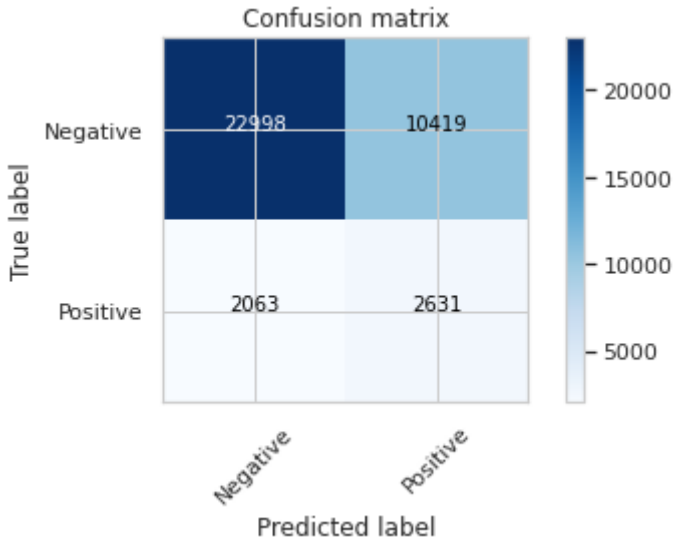
	precision	recall	f1-score	support
Negative	0.88	1.00	0.93	33417
Positive	0.00	0.00	0.00	4694
accuracy			0.88	38111
macro avg	0.44	0.50	0.47	38111
weighted avg	0.77	0.88	0.82	38111

Το under-sampling των αρνητικών απαντήσεων (Κλάση 0) δεν επηρέασε σε κανένα βαθμό το μοντέλο.

Μοντέλο #3

Το μοντέλο αυτό χρησιμοποιεί την ιδιά ακριβώς αρχιτεκτονική με το μοντέλο 1 άλλα στα δεδομένα εκπαίδευσης του έχει πραγματοποιηθεί (επιπλέον) under-sampling (από 270.884 δείγματα σε 100.000).

Confusion Matrix:



Classification Report

	precision	recall	f1-score	support
Negative	0.92	0.69	0.79	33417
Positive	0.20	0.56	0.30	4694
accuracy			0.67	38111
macro avg	0.56	0.62	0.54	38111
weighted avg	0.83	0.67	0.73	38111

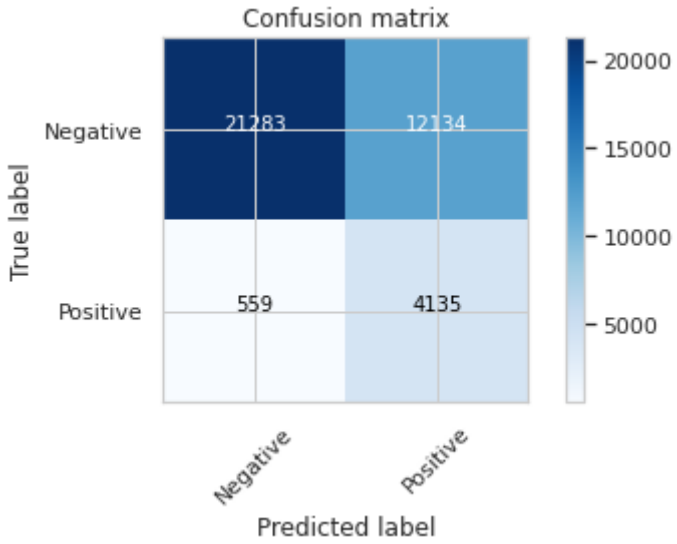
Για πρώτη φορά το μοντέλο κάνει προβλέψεις και για θετικές απαντήσεις πελατών και μάλιστα κατάφερε να

εντοπίσει πάνω από τους μίσους πελάτες που θα απαντήσουν θετικά βελτιώνοντας το Macro Average f1 score σε 0.54 από 0.47. Οι προβλέψεις αυτές όμως συνοδεύονται από ένα κόστος στην μορφή των 10.419 λανθασμένων θετικών προβλέψεων που έγιναν για πελάτες που θα έδιναν αρνητική απάντηση ρίχνοντας κατά αυτό τον τρόπο το accuracy σε 0.67.

Μοντέλο #4

Το μοντέλο αυτό είναι πανομοιότυπο με το προηγούμενο με μονή διαφορά ότι έχει πραγματοποιηθεί over-sampling στα δείγματα της Κλάσης 1 (Θετικές απαντήσεις) από 37.814 σε 60.000.

Confusion Matrix:



Classification Report

	precision	recall	f1-score	support
Negative	0.97	0.64	0.77	33417
Positive	0.25	0.88	0.39	4694
accuracy			0.67	38111
macro avg	0.61	0.76	0.58	38111
weighted avg	0.89	0.67	0.72	38111

Το accuracy του μοντέλου παραμένει σταθερό στο 67% όμως βλέπουμε περαιτέρω βελτίωση της μετρικής Macro Average f1 score σε 0.58. Το μοντέλο επίσης βρίσκει όλες σχεδόν τους πελάτες που δίνουν θετική απάντηση.

Ο παρακάτω πίνακας συνοψίζει τα αποτελέσματα των νευρικών δικτυών που δημιουργήθηκαν έως τώρα:

	Μοντέλο #1	**Μοντέλο #2**	**Μοντέλο #3**	**Μοντέλο #4**
Accuracy	0.88	0.88	0.67	0.67
Macro Average f1 score	0.47	0.47	0.54	0.58
Total of Positive Responses found	2	0	2631	4135

Cost-Sensitive Neural Networks

Η χρήση του Resampling βελτίωσε σε μεγάλο βαθμό την απόδοση των μοντέλων. Αποτελεί όμως μια γενική μέθοδος επίλυσης προβλημάτων Μηχανικής Μάθησης που χρησιμοποιούν Imbalanced datasets ενώ για την περίπτωση των Νευρωνικών Δικτυών υπάρχει ειδική μέθοδος αντιμετώπισης αυτού του φαινομένου.

Κάνοντας χρήση των Cost-Sensitive Νευρωνικών Δικτυών (γνωστά και ως Weighted Neural Networks) δίνεται η δυνατότητα να ορίζουμε βάρη σε κάθε κλάση επιτρέποντας στο δίκτυο να αντιμετωπίζει με μεγαλύτερη βαρύτητα παραδείγματα από το Minority Class σε σχέση με το Majority Class.

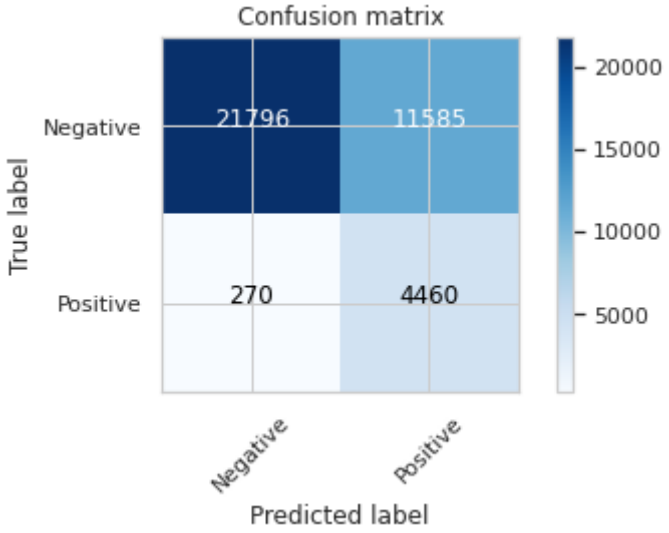
Μοντέλο #5

Μια καλή τακτική όσο αναφορά τα βάρη που θα ανατεθούν σε κάθε κλάση είναι να εφαρμόζεται το αντίστροφο της αναλογίας των δυο κλάσεων. Για το συγκεκριμένο training dataset η αναλογία αναμεσά σε Minority και Majority class είναι ίση με 46.710 : 334.399 η οποία απλοποιημένη αντιστοιχεί σε αναλογία 14:100. Το βάρος επομένως της Κλάσης 1 των θετικών απαντήσεων στο μοντέλο αυτό θα είναι 100 ενώ για τη Κλάση 0 θα είναι 14.

Επιπλέον στο μοντέλο αυτό εφαρμόζεται διαφορετική αρχιτεκτονική νευρωνικού δικτύου σε σχέση με τα

προηγούμενα. Αποτελείται από 1 Hidden Layer με 10 ReLU νευρώνες το οποίο καταλήγει σε ένα Sigmoid νευρώνα. Τα αρχικά βάρη του δικτύου επιλέγονται από τον He Uniform Initializer, η ιδιά Early Stopping τεχνική με τα προηγούμενα δίκτυα εφαρμόζεται και εδώ και τέλος, ως optimizer χρησιμοποιείται η τεχνική Stochastic Gradient Descent.

Confusion Matrix:



Classification Report

	precision	recall	f1-score	support
Negative	0.99	0.65	0.79	33381
Positive	0.28	0.94	0.43	4730
accuracy			0.69	38111
macro avg	0.63	0.80	0.61	38111
weighted avg	0.90	0.69	0.74	38111

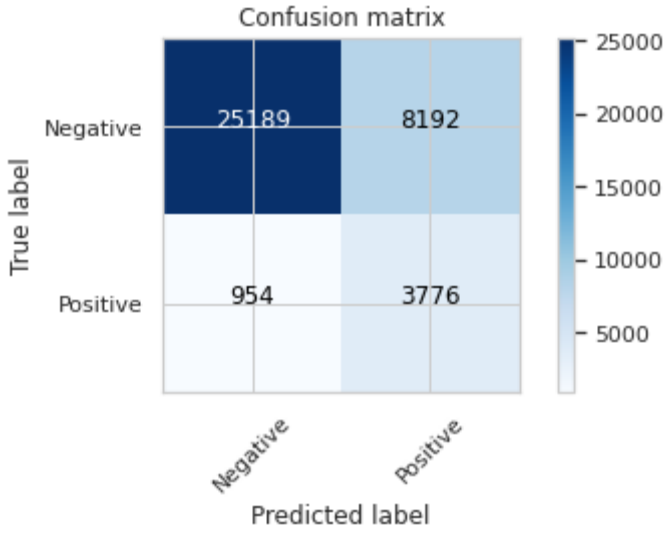
Τα Cost-Sensitive νευρωνικά δίκτυα φαίνεται να είναι η λύση για το πρόβλημα. Στα αποτελέσματα φαίνεται η ταυτόχρονη βελτίωση και του Accuracy και του Macro Average f1 score. Το δίκτυο εντοπίζει σχεδόν όλους

τους πελάτες που είναι θετικοί σε προσφορά ασφάλειας αυτοκίνητου.

Μοντέλο #6

Το μοντέλο είναι πανομοιότυπο με το προηγούμενο άλλα χρησιμοποιεί διαφορετικό ratio για τα βάρη των κλάσεων (28 για την κλάση 0 και 100 για την κλάση 1), έτσι ώστε να μειωθεί η μεγάλη προτεραιότητα που δίνεται στην κλάση 1, με σκοπό την μεγιστοποίηση του Macro Average f1 score.

Confusion Matrix:



Classification Report

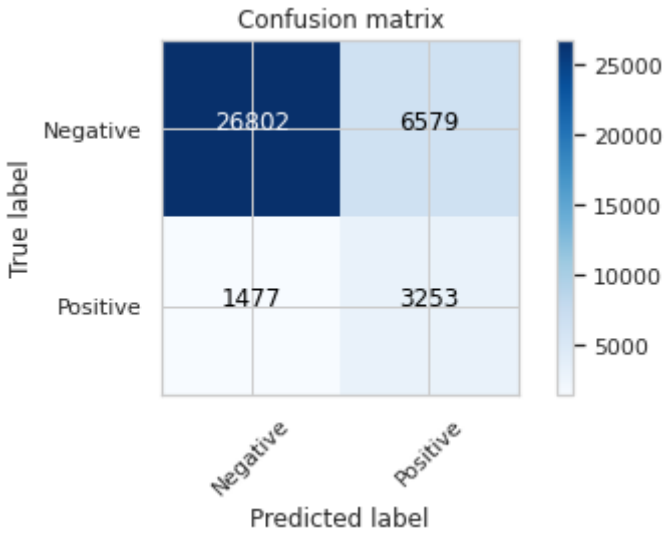
	precision	recall	f1-score	support
Negative	0.96	0.75	0.85	33381
Positive	0.32	0.80	0.45	4730
accuracy			0.76	38111
macro avg	0.64	0.78	0.65	38111
weighted avg	0.88	0.76	0.80	38111

Το μικρό αυτό adjustment στα βάρη που δίνονται σε κάθε κλάση ήταν αρκετό ώστε να γίνει ταυτόχρονη αύξηση και στο accuracy (0.76 από 0.69) και στο Macro Average f1 score (0.65 από 0.61).

Μοντέλο #7

Χρησιμοποιείται η ιδιά αρχιτεκτονική για το δίκτυο, με μια μικρή προσαρμογή στα βάρη κάθε κλάσης (35 για την κλάση 0 και 100 για την κλάση 1).

Confusion Matrix:



Classification Report

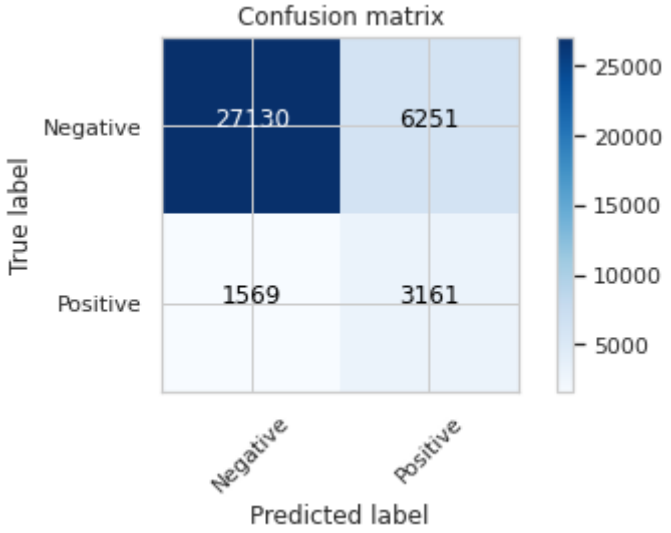
	precision	recall	f1-score	support
Negative	0.95	0.80	0.87	33381
Positive	0.33	0.69	0.45	4730
accuracy			0.79	38111
macro avg	0.64	0.75	0.66	38111
weighted avg	0.87	0.79	0.82	38111

Επιπλέον βελτίωση (αν και μικρή) στο μοντέλο λόγω αυτής της μικρής προσαρμογής.

Μοντέλο #8

Στο σημείο αυτό διατηρείται η αναλογία για τα βάρη των κλάσεων που πέτυχε την καλύτερη απόδοση (35:100) και δοκιμάζονται διάφορες αρχιτεκτονικές για τα μοντέλα. Για το συγκεκριμένο μοντέλο, χρησιμοποιούνται 2 Hidden Layers των 16 και 32 ReLU νευρώνων και ένας sigmoid νευρώνας στο Output Layer. Επίσης χρησιμοποιείται ο Adam optimizer και Early Stopping.

Confusion Matrix:



Classification Report

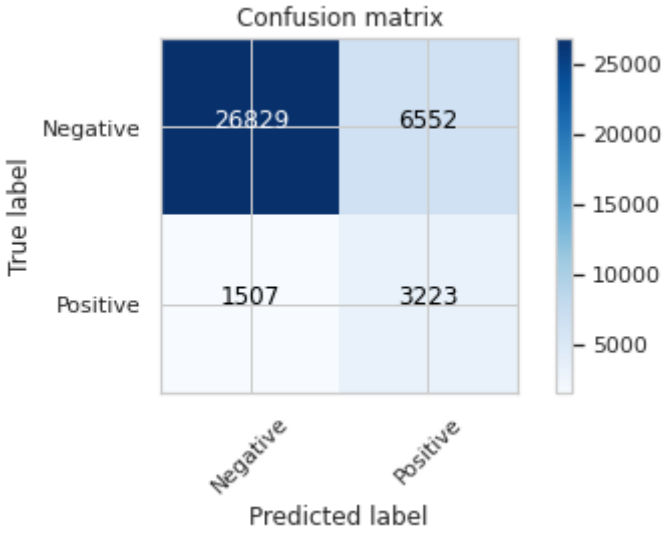
	precision	recall	f1-score	support
Negative	0.95	0.81	0.87	33381
Positive	0.34	0.67	0.45	4730
accuracy			0.79	38111
macro avg	0.64	0.74	0.66	38111
weighted avg	0.87	0.79	0.82	38111

Η διαφορετική αρχιτεκτονική δεν δείχνει να επηρεάζει τα αποτελέσματα με κάποιο τρόπο.

Μοντέλο #9

Το νευρωνικό δίκτυο αυτό χρησιμοποιεί τα ιδιά βάρη για τις κλάσεις άλλα διαφορετική αρχιτεκτονική, η οποία περιλαμβάνει ένα Hidden Layer με 60 ReLU νευρώνες. Μια άλλη βασική διαφορά είναι πως αντί για Early Stopping γίνεται χρήση ενός διαφορετικού είδους callback, το λεγόμενο Reduce_Learning_Rate. Η τεχνική αυτή υλοποιείται με παρακολούθηση του δίκτυο και πιο συγκεκριμένα, της μετρικής Validation AUC, και σε περίπτωση που δεν εντοπιστεί κάποια βελτίωση σε αυτή για 5 εποχές τότε μειώνεται ο ρυθμός μάθησης.

Confusion Matrix:



Classification Report

	precision	recall	f1-score	support
Negative	0.95	0.80	0.87	33381
Positive	0.33	0.68	0.44	4730
accuracy			0.79	38111
macro avg	0.64	0.74	0.66	38111
weighted avg	0.87	0.79	0.82	38111

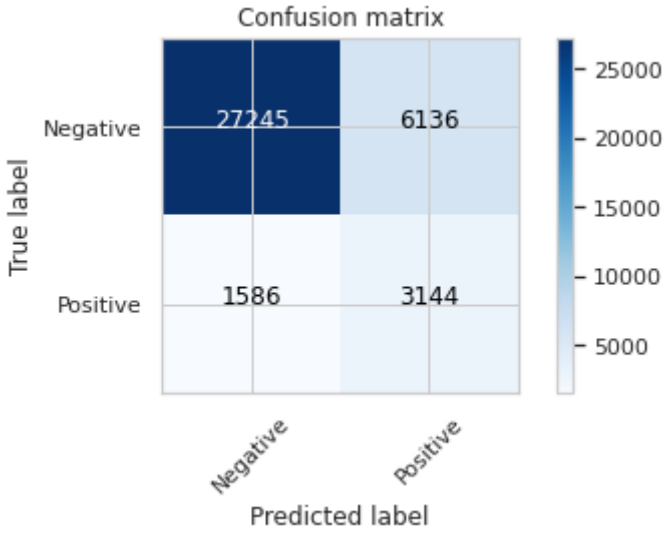
Τα αποτελέσματα παραμένουν σταθερά.

Μοντέλο #10

Όμοιο μοντέλο με το προηγούμενο με λίγες διάφορες:

- 2 Hidden Layers με 60 και 30 ReLU νευρώνες
- RMSprop optimizer με 0.1 learning rate

Confusion Matrix:



Classification Report

	precision	recall	f1-score	support
Negative	0.94	0.82	0.88	33381
Positive	0.34	0.66	0.45	4730
accuracy			0.80	38111
macro avg	0.64	0.74	0.66	38111
weighted avg	0.87	0.80	0.82	38111

Αν και με μικρή διαφορά, το νευρωνικού δίκτυο αυτό έχει την καλύτερη απόδοση. Οι παρακάτω πίνακες συνοψίζουν τα αποτελέσματα:

	Μοντέλο #5	Μοντέλο #6	Μοντέλο #7	Μοντέλο #8
Accuracy	0.69	0.76	0.79	0.79
Macro Average f1 score	0.61	0.65	0.66	0.66
Total of Positive Responses found	4.460	3.776	3.253	3.161

	Μοντέλο #9	Μοντέλο #10
Accuracy	0.79	0.80
Macro Average f1 score	0.66	0.66
Total of Positive Responses found	3.223	3.144

Conclusions

Η παρούσα εργασία κατάφερε να αντιμετωπίσει σε μεγάλο βαθμό την ανισορροπία που χαρακτηρίζει το dataset που ανέλαβε και πέτυχε την δημιουργία χρήσιμων μοντέλων χάρη στην τεχνική των Cost-Sensitive νευρωνικών δικτυών.

Σε περίπτωση που το Business Goal της μελέτης αυτής ήταν ένα μοντέλο που προβλέπει όσες περισσότερες θετικές απαντήσεις γίνεται σε μια λίστα πελατών τότε το ιδανικό μοντέλο είναι το 5ο, το οποίο εντόπισε 4.460 θετικές απαντήσεις από τις 4.730.

Αν όμως στοχεύουμε στην δημιουργία ενός αντικειμενικού μοντέλου που διαχωρίζει αποτελεσματικά τις 2 κλάσεις τότε το καλύτερο μοντέλο είναι το 10ο, το οποίο πέτυχε την υψηλότερη τιμή Average f1 score (0.66) διατηρώντας παράλληλα ένα Accuracy της τάξης του 80%

Σημείωση:

Τα δυο καλυτέρα μοντέλα έχουν αποθηκευτεί, όπως και τα train, test και validation datasets υστέρα από την σχετική επεξεργασία και θα βρίσκονται στο Google Drive. Παράλληλα, το Google Colab Notebook, το οποίο και αυτό θα βρίσκεται στο drive περιέχει ολόκληρο τον κώδικα της εργασίας καθώς και μια αναλυτική επεξήγηση της διαδικασίας.