Internship at LInC Summer 2020
Final Report

Intern
Στυλιανός Ηροδότου

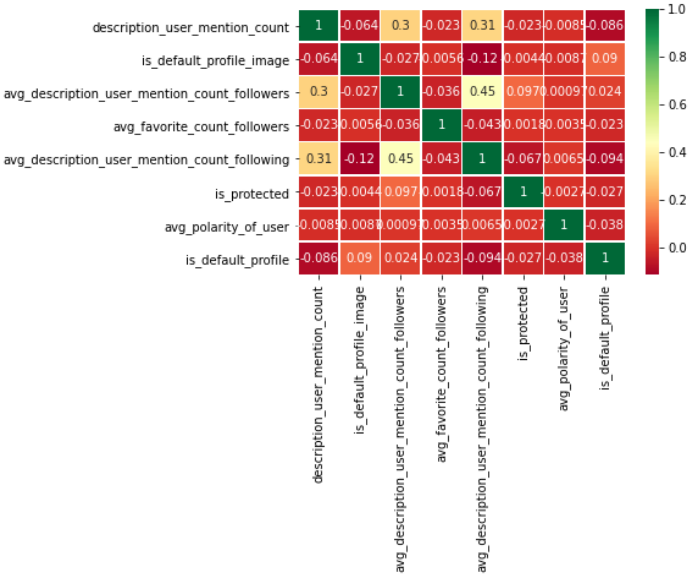
Εργοδότες
Δρ. Γεώργιος Πάλλης και
Δρ. Μάριος Δικαιάκος

Internship Supervisor's
Δημοσθένης Στεφανίδης

Abstract

Ο σκοπός του Project του Internship ήταν η περαιτέρω κατανόηση του πώς το κοινωνικό μας δίκτυο επηρεάζει την επαγγελματική μας κατάρτιση. Συγκεκριμένα σε αυτό το project εξέτασα πώς διάφορα χαρακτηριστικά του Δικτύου κάποιου entrepreneur στο Twitter μπορεί να επηρεάσουν κατά πόσο θα πάρει χρηματοδότηση ή όχι. Για να γίνει αυτό χρειάστηκε να συλλέξω πολλά δεδομένα διαφόρων entrepreneurs και του δικτύου τους. Χρειάστηκε να συλλέξω δεδομένα για τους ίδιους καθώς και για τους followers και followings τους. Στη συνέχεια, επεξεργάστηκα τα δεδομένα σε μορφή κειμένου με τη χρήση του natural language toolkit (nltk), ώστε να βγάλω διάφορα συμπεράσματα όπως sentiment, polarity, subjectivity hashtag usage και user mentions μεταξύ άλλων. Έπειτα, έγινε περεταίρω επεξεργασία, φιλτράρισμα και συνδυασμός των δεδομένων για να δημιουργηθούν χρήσιμες πλέον πληροφορίες. Στο τέλος της προ επεξεργασίας, κατέληξα με ένα Dataset με 78 μεταβλητές.

Συνέχισα δημιουργώντας μερικά Γραφήματα για να καταλάβω περεταίρω τη δομή των δεδομένων μου, και στη συνέχεια έσμιξα μεταβλητές με εξαιρετικά ψηλό linear correlation.



Correlation heatmap (for 8 most important features for saving space)

Μετά με τη χρήση 4 τεχνικών feature selection καταλήγω στις 24 πιο σημαντικές μεταβλητές και βρίσκω τις βέλτιστές hyperparameters για 7 μοντέλα binary classification. Στη συνέχεια, συγκρίνω τα μοντέλα αυτά με τη χρίση Stratified KFold cross validation και επιλέγω το καλύτερο το οποίο ελέγχω στο test set.

Ακολουθώντας τα πιο πάνω βήματα κατάφερα να δημιουργήσω ένα μοντέλο το οποίο με είσοδο τις πληροφορίες του δικτύου ενός entrepreneur στο Twitter βρίσκει κατά πόσο θα πάρει χρηματοδότηση ή όχι με classification report weighted average F1 score 0.71167

Motivation

Στις μέρες μας, είναι προφανές πως τα μέσα κοινωνικής δικτυώσεις μπορούν να επηρεάσουν την επαγγελματική μας κατάρτιση καθώς πρακτικές όπως social media look up πριν μια συνέντευξη και social media marketing έχουν πλέον καθιερωθεί. Πώς όμως το κοινωνικό δίκτυο κάποιου επηρεάζει τις πιθανότητες του να πάρει χρηματοδότηση;

Είναι εξαιρετικά ενδιαφέρον πως μερικές φαινομενικά ασήμαντες πράξεις όπως ο τρόπος που γράφει κάποιος ένα status και ποιους ακολουθεί στο προσωπικό του Λογαριασμό, μπορούν να επηρεάσουν θετικά ή αρνητικά τη επαγγελματική του προοπτική. Για το λόγο αυτό ήθελα να ανακαλύψω με ποιους τρόπους μπορώ να χρησιμοποιήσω το Twitter για να βελτιστοποιήσω την προοπτική αυτή.

Challenges of Project

Το Twitter θέτει περιορισμούς για πόσα δεδομένα μπορεί κάποιος χρήστης να συλλέξει σε συγκεκριμένα χρονικά πλαίσια. Αυτός είναι και ο λόγος που έσπερνε τόσο χρόνο η συλλογή δεδομένων και κα επέκταση γιατί το δείγμα είναι πολύ μικρό.

Επίσης καθώς το κοινωνικό δίκτυο κάποιου δεν έχει τον καθοριστικό ρόλο κατά πόσο θα πάρει χρηματοδότηση ή όχι, με το να περιορίζω της πληροφορίες που χρησιμοποιεί ο αλγόριθμος στο κοινωνικό του δίκτυο περιορίζω την προοπτική του αλγορίθμου.

Implementation Challenges

Μια απρόσμενη δυσκολία που είχα ήταν το πρόβλημα που είχα με το VM που με κράτησε περίπου μια εβδομάδα πίσω όσο αφορά τη συλλογή δεδομένων για αυτό δεν πρόλαβα να συλλέξω όσα δεδομένα θα ήθελα και το τελικό μου δείγμα ήταν 850 αντί 1000 και περιορίστηκα σε συγκεκριμένο αριθμό Followers και Followings για κάθε entrepreneur.

Τέλος, μια μικρή δυσκολία ήταν να δουλέψω από απόσταση. Ιδανικά θα ήθελα να είχα δουλέψει πιο κοντά με τους προϊστάμενους μου.

Methodology Overview

Η μεθοδολογία που ακολούθησα χωρίζεται ως εξής:

1. Data Collection
2. Feature Selection
3. Find and Train Optimal Model

Data Collection

Το πρώτο βήμα για την δημιουργία του μοντέλου ήταν η εύρεση δεδομένων από το Twitter. Για να γίνει αυτό χρησιμοποίησα το Tweepy που είναι μια βιβλιοθήκη για τη Python για ευκολότερη χρήση του Twitter API. Η αποθήκευση των δεδομένων έγινε σε MongoDB.

Αρχικά κατέβασα μερικά δεδομένα τύπου χρήστη και τύπου tweet, έμαθα τη δομή τους και πήρα μια ιδέα ποιες πληροφορίες θα μπορούσα να χρησιμοποιήσω. Στη συνέχεια, δημιούργησα μεθόδους με τις οποίες βρήκα πόσο χρόνο χρειάζομαι για τη συλλογή διαφόρων ειδών δεδομένων για διάφορα μεγέθη δείγματος. Μαζί με το Δημοσθένη, αποφασίσαμε πώς λόγο των μικρών χρονικών πλαισίων που είχα για το internship το τελικό δείγμα θα είναι μόνο 1000 entrepreneurs.

Όταν αποφασίστηκε αυτό, δημιούργησα μεθόδους με τις οποίες θα έπαιρνα τα δεδομένα από το Twitter και θα τα επεξεργαζόμουν.

Δυστυχώς, η διαδικασία δεν κύλισε όσο ομαλά όσο θα ήθελα, και λόγο της αδυναμίας μου να συλλέξω δεδομένα τη μια εβδομάδα από όσες μέρες περίμενα πως θα έχω, το τελικό δείγμα περιορίστηκε τελικά σε 850 entrepreneurs ο κάθε ένας με ένα δίκτυο 400 ατόμων, 200 followers και 200 followings, και ο κάθε χρήστης από το δίκτυο είχε μέγιστο όριο 800 tweets.

Feature Selection

Αρχικά ο τρόπος που θα αντιμετώπισα missing values ήταν να τις αντικαταστήσω με το median αλλά με τον τρόπο που έκανα την προ επεξεργασία δεν υπήρχαν missing values. Επίσης, ένωσα μεταβλητές που ήταν Highly linearly corelated.

Στη συνέχεια χρησιμοποίησα τις πιο κάτω τεχνικές για την μείωση των μεταβλητών:

1. Removing features with low variance:
   It removes all features whose variance doesn't meet some threshold. By default, it removes all zero-variance features, i.e. features that have the same value in all samples.
2. Tree-based feature selection
   Finds features that have zero importance according to a gradient boosting machine learning model. Such models are tree-based machine learning models, that can find feature importance. In a tree-based model, the features with zero importance are not used to split any nodes, and so they can be removed without affecting model performance.
3. Univariate feature selection
   selecting the best features based on univariate statistical tests, removes all but the k highest scoring features
4. Recursive feature elimination with cross-validation
   A recursive feature elimination example with automatic tuning of the number of features selected with cross-validation.

Find and Train Optimal Model

Στη συνέχεια, δημιούργησα test set και train set using stratified sampling για να αποφύγουμε το sampling bias και βρήκα κοντά σε βέλτιστες hyperparameteres των πιο κάτω μοντέλων με RandomizedSearchCV:

- Stochastic Gradient Descent
- Logistic Regression
- k-Nearest Neighbors
- Decision Trees
- Support Vector Machine
- Naive Bayes
- Random Forest

Έπειτα, συγκρίνω τα μοντέλα χρησιμοποιώντας stratified k fold cross validation και επιλέγω το καλύτερο.

Findings:

Έκρινα πώς η ακρίβεια (Accuracy) δεν ήταν καλός τρόπος αξιολόγησης του μοντέλου λόγω του Accuracy Paradox που με λίγα λόγια εξήγα πώς το Accuracy δουλεύει καλά μόνο όταν και τα δύο πιθανά αποτελέσματα(στη περίπτωση μας κάποιος να πάρει χρηματοδότηση ή όχι) είναι ίσα στο δείγμα μας. Αυτή η έλλειψή ισορροπίας κάνει το Accuracy μη αξιόπιστο μετρικό. Ο τρόπος που αξιολογώ το μοντέλο είναι με F1-score. F1-score είναι το weighted average του Precision και Recall τα οποία λαμβάνουν υπόψη τους τα false negatives και τα false positives. Έκρινα πως αυτός ο τρόπος αξιολογήσεις ενός binary classifier είναι βέλτιστος για τους σκοπούς του Project.

Το καλύτερο μοντέλο από το cross validation ήταν το Support Vector Machine και το αποτέλεσμα του classification report για weighted average στο τεστ σετ είναι:

```
recall: 0.65465342955
precision 0.77955488444
f1_score 0.711665486307672
support 130
```

Ακολουθεί πίνακας με τα αποτελέσματα όλων των μοντέλων στο Τεστ Σετ

Table with results

Logistic Regression model	recall: 0.49955488444 precision 1.0 f1_score 0.6662708909471571 support 130
K Neighbours Classifier model	recall: 0.5 precision 0.5078106224465273 f1_score 0.4988455390444616 support 130
Gaussian Naïve Bayes model	recall: 0.5461538461538461 precision 0.9576396206533193 f1_score 0.6955989284347494 support 130
Decision Tree Classifier model	recall: 0.5538461538461539 precision 0.5538461538461539 f1_score 0.5538461538461539 support 130
Support Vector Classification model	recall: 0.65465342955 precision 0.77955488444 f1_score 0.711665486307672 support 130
SGD Classifier model	recall: 0. 6169230769230769 precision 0. 61955488444 f1_score 0.6182361798185135 support 130
Random Forest model	recall: 0.6076923076923076 precision 0.6070582145564122 f1_score 0.6073394072194671 support 130

Conclusion and Future Work

Εν κατακλείδι, δεδομένου ότι το κοινωνικό δίκτυο κάποιου δεν έχει τον καθοριστικό ρόλο κατά πόσο θα πάρει χρηματοδότηση ή όχι αλλά το προϊόν που παράγει, χαίρομαι που δημιουργήθηκε ένα μοντέλο με ικανοποιητική ακρίβεια αλλά πιστεύω πως εάν υπήρχε περισσότερος χρόνος υπάρχουν πολλά που θα μπορούσαν να βελτιωθούν για να αυξήσουν περεταίρω το F1-score του αλγορίθμου.

Το πιο προφανές είναι η ποσότητα δεδομένων. Το τελικό μου δείγμα ήταν εξαιρετικά μικρό λόγο του λίγου χρόνου που είχα και αφού η μέθοδοι είναι ήδη έτοιμες και το Twitter δεν έχει κάποιο όριο στο πόσα δεδομένα μπορούμε να συλλέξουμε in the long run, θα ήταν το πρώτο που θα έκανα.

Επίσης θα ήθελα να επεκτείνω τα δεδομένα μου.

1. Θα ήθελα να συμπεριλάβω την επεξεργασία φωτογραφιών κάτι που είναι εξαιρετικά σημαντικό στα μέσα κοινωνικής δικτυώσεις. Είναι γνωστό ότι συμπεριφορές όπως το χαμόγελο αυξάνουν τις πιθανότητες να γίνει κάποιος αρεστός, και κατά συνέπεια δίνουμε περισσότερες ευκαιρίες σε άτομα μου μας είναι αρεστά. Με το πιο πάνω συλλογισμό δεν μπορώ παρά να αναρωτηθώ με ποιο τρόπο θα επηρέαζέ τις πιθανότητές κάποιου να πάρει χρηματοδότηση. Υπάρχει έτοιμο pipeline που βρίσκει πόσο άτομα είναι σε μια φωτογραφία, αναγνωρίζει ποιοι είναι μα βάση ποια άτομα είναι στη βάση δεδομένων και αναγνωρίζει το συναίσθημα τους. Επίσης βρήκα αλγόριθμους που ανιχνεύουν filters και τροποποίηση εικόνας (photoshop) που επίσης θα μπορούσε να είναι ένα μετρικό.
2. Ένα βήμα περεταίρω θα μπορούσα να επεξεργαστώ Βίντεο και live streaming videos.
3. Θα μπορούσα να συμπεριλάβω διάφορα δεδομένα που διάλεξα να μην συμπεριλάβω λόγο έλλειψης χρόνου. Συγκεκριμένα θα μπορούσα να χρησιμοποιήσω διαφορά δεδομένα για να βρω περίπου τη θέση κάποιου χρήστη για να μετρήσω το εύρος του δικτιού, δηλαδή ένα κάποιος χρήστης έχει φίλους μόνο στην χώρα/ πόλη του, δηλαδή το δίκτυο του είναι τοπικό, ή εάν έχει φίλους σε πολλές χώρες

Τέλος, θα μπορούσα να χρησιμοποιήσω περισσότερα Μοντέλα και αντί για RandomizedSearchCV να χρησιμοποιούσα Grid Search CV για τις βέλτιστες hyperparameters.

Κλείνοντας, θα ήθελα να ευχαριστήσω τον Κύριο Πάλλη για την ευκαιρία να δουλέψω σε ένα τόσο ενδιαφέρον Project και τον Δημοσθένη Στεφανίδη για την εξαιρετική καθοδήγήση του καθ'όλη τη διάρκεια του Project, μερικές φορές ακόμα και τα σαββατοκύριακα αλλά και αργά τη νύχτα.

Ευχαριστώ.

Κώδικας:

GitHub: https://github.com/StylianosHerodotou/How-different-is-your-Twitter-network-from-you

Jupiter Notebook: 10.16.3.55:6670/?token=f9fa30e478adf7ffd0198f32814901db2b4cdf64c8f78ce9