מתודולוגיות עבודה ב-DS

תהליך עבודה בפרויקט Data Science

תוכן עניינים

- הגדרת הבעיה
- הכנת המידע
- הרצת אלגוריתם ראשוני
- שיפור התוצאות
- הצגת התוצאות

Home Credit Default Risk

- רקע
  - Many people struggle to get loans due to insufficient or non-existent credit histories. And, unfortunately, this population is often taken advantage of by untrustworthy lenders.
  - Home Credit strives to broaden financial inclusion for the unbanked population by providing a positive and safe borrowing experience. In order to make sure this underserved population has a positive loan experience, Home Credit makes use of a variety of alternative data--including telco and transactional information--to predict their clients' repayment abilities.

Home Credit Default Risk

- רקע
  - While Home Credit is currently using various statistical and machine learning methods to make these predictions, they're challenging Kagglers to help them unlock the full potential of their data. Doing so will ensure that clients capable of repayment are not rejected and that loans are given with a principal, maturity, and repayment calendar that will empower their clients to be successful.
  - For each SK_ID_CURR in the test set, you must predict a probability for the TARGET variable.
  - Submissions are evaluated on area under the ROC curve between the predicted probability and the observed target.

הגדרת הבעיה

1. ניסוח פשוט - יש לסווג / לתת הסתברות האם משק בית יוכל לעמוד בתשלומי אשראי
2. ניסוח רשמי

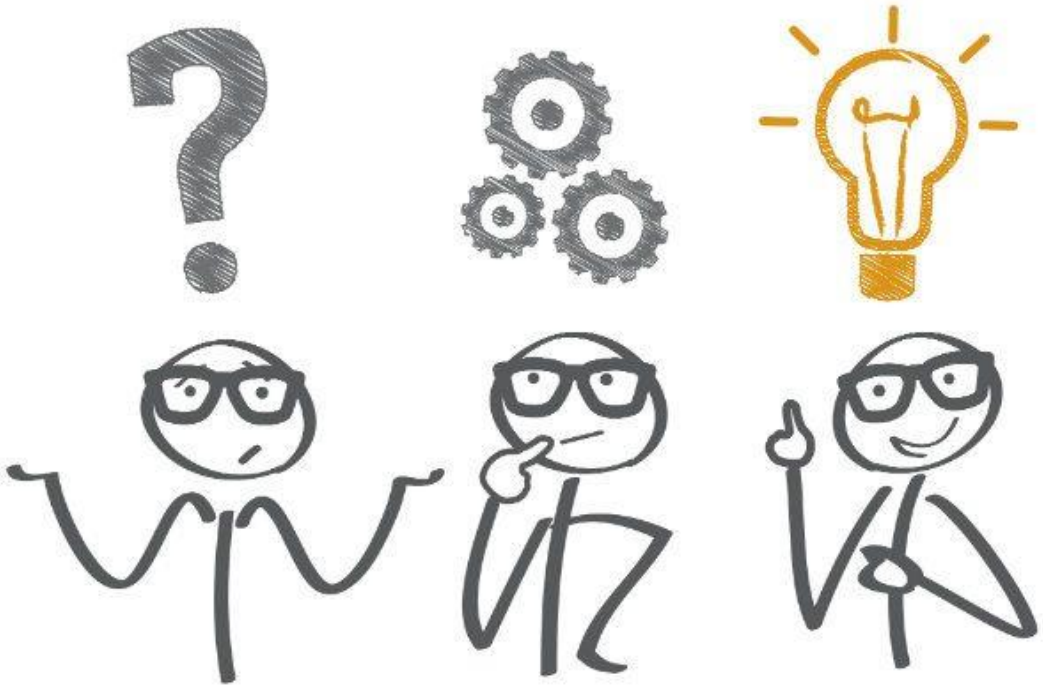
- בהינתן קבוצה של משקי בית ויכולת ההחזר שלהם יש לסווג / לתת הסתברות עבור יכולת ההחזר של משק בית חדש

הגדרת הבעיה

3. הנחות לפתירת הבעיה

- מהי ההגדרה למשק בית שלא עמד בהחזרי אשראי? - מומלץ לבצע מחקר בנושא
  - מידע רלוונטי לפתרון הבעיה
    - AMT_INCOME_TOTAL - Income of the client
  - מידע שלא רלוונטי למודל
    - WEEKDAY_APPR_PROCESS_START - On which day of the week did the client apply for the loan
  - אמינות הנתונים
    - REGION_RATING_CLIENT - Our rating of the region where client lives (1,2,3)
  - עדכניות הנתונים
  - כמות הדוגמאות החיוביות והשליליות שצריכות להכנס למודל
  - מציאת בעיות דומות

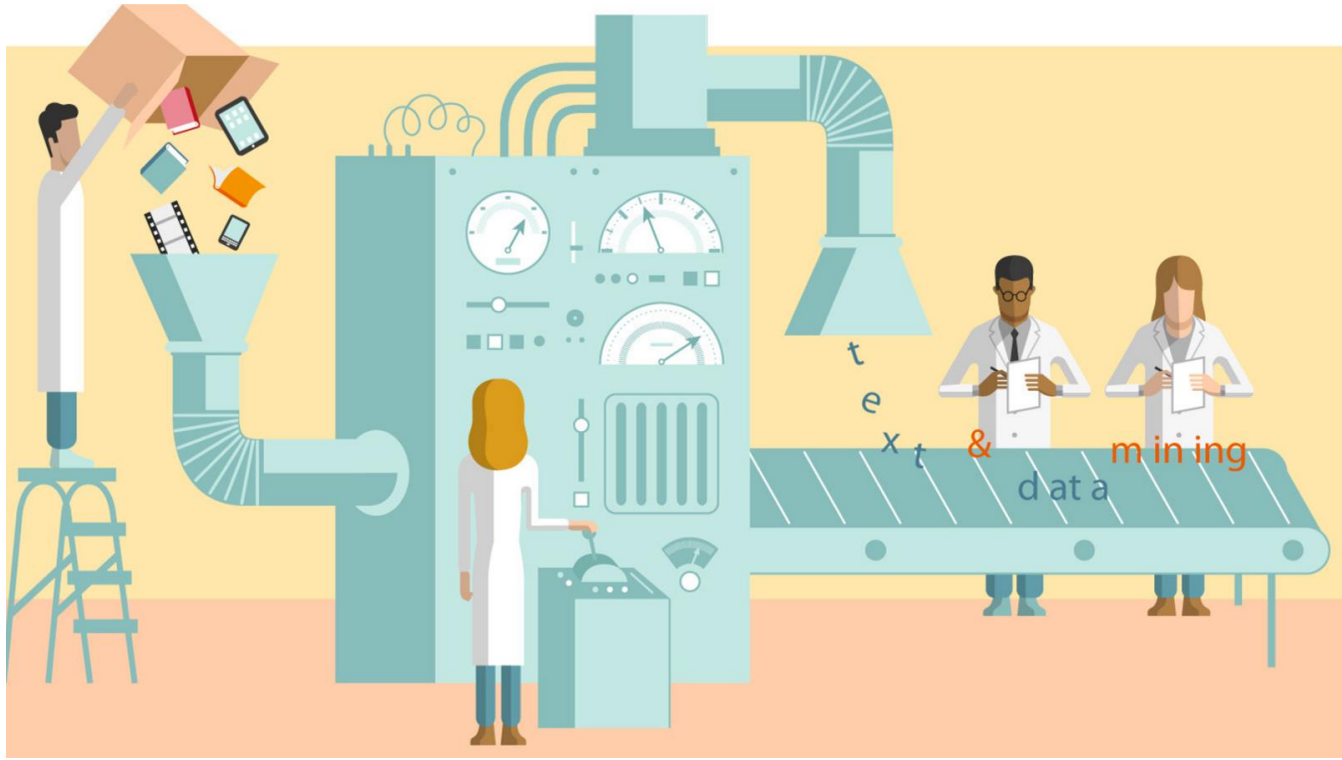
הגדרת הבעיה



5. איך עלי לפתור את הבעיה? - איך הייתי פותר את הבעיה ידנית

- איזה מידע יש לאסוף
  - דוגמאות - מהי ההגדרה למשק בית שלא עמד בהחזרי אשראי? , גורמים לשוני בין דוגמאות - לדוג' אירועים מיוחדים / משבר כלכלי
  - מאפיינים - איך הייתי מתחקר את הנתונים?
- איך לאסוף את המידע הדרוש
- איזה עיבוד מידע יבוצע

הכנת המידע



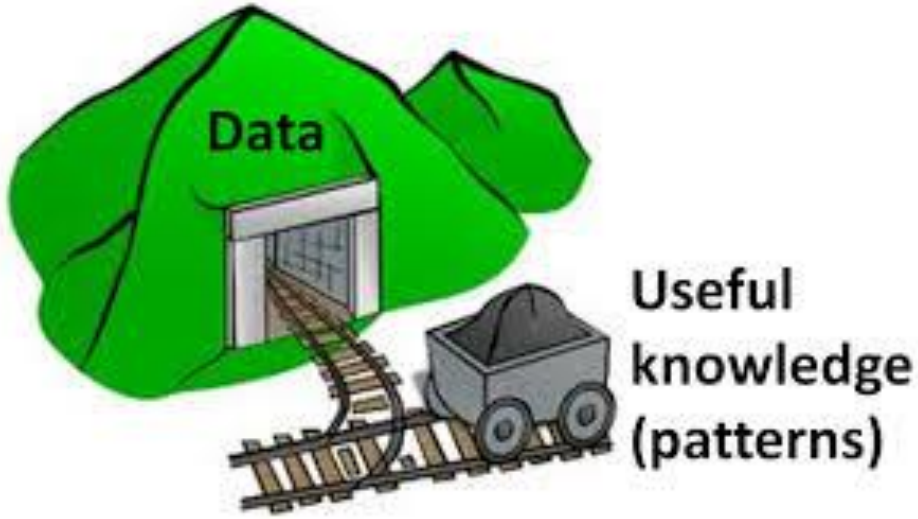
1. תהליך הכנת המידע*
   1. בחירת המידע
   - ביצוע עיבוד מקדים **Feature Engineering**
   1. transform data
   2. בדיקת הנתונים (groups , outliers)
- במידת הצורך ניתן להציג ויזואלית את המידע (Data Exploration)
- קבוצות נוצרות מ-variance בין הדוגמאות

הכנת המידע

- 1 - בחירת המידע
  1. בחירת הנתונים הרלוונטיים לפתרון הבעיה
     1. דוגמאות רלוונטיות - בעלי variance נמוך וללא אנומליות
     2. מאפיינים רלוונטיים - ע"פ מומחיות התוכן, איך הייתי פותר את הבעיה ידנית?
  2. איזה נתונים ניתן להביא
  3. איזה מידע לא ניתן להביא מה-DB - האם ניתן לבצע סימולציה (לדוג' לימוד מאפיינים)

הכנת המידע

- 2 - ביצוע עיבוד מקדים
  1. Sampling - במידה ואוספים כמות מוגבלת של דוגמאות יש לוודא שהם מהוות מדגם מייצג
  2. Formatting - איחוד המידע לפורמט אחיד
  3. Cleaning - ניקוי המידע (טיפול בערכים חסרים)



Feature Engineering

- Feature Engineering עוסק בייצוג נכון של הבעיה
- ב-Feature Engineering ניתן מענה על השאלה, איך ניתן להציג את המידע בצורה הטובה ביותר בכדי לפתור את הבעיה?
- תהליך ה-Feature Engineering כולל, הערכת חשיבות המאפיינים וחילוצם
- תהליך הערכת חשיבות המאפיינים יכול להתבצע באופן ידני או אוטומטי

הכנת המידע

- 3 - שינוי המידע (transform data)
  - Scaling - הצגת נתונים המגיעים בגדלים שונים בפורמט מספרי אחיד (Z-score - חשוב ב-NN ולא בעצי החלטה, time scaling, אחוזים - במידה והמידה מוטה)
    בין 0 ל-1 או סביב ה-0 (feature scaling or standart score)
  - Decomposition - פירוק מאפיינים מורכבים לרכיביהם (כגון פירוק שדה תאריך ליום ושעה)
  - Aggregation - איחוד של מאפיינים (לדוג' איחוד קבוצות גיל)
  - Categorical vs Continuous

הכנת המידע

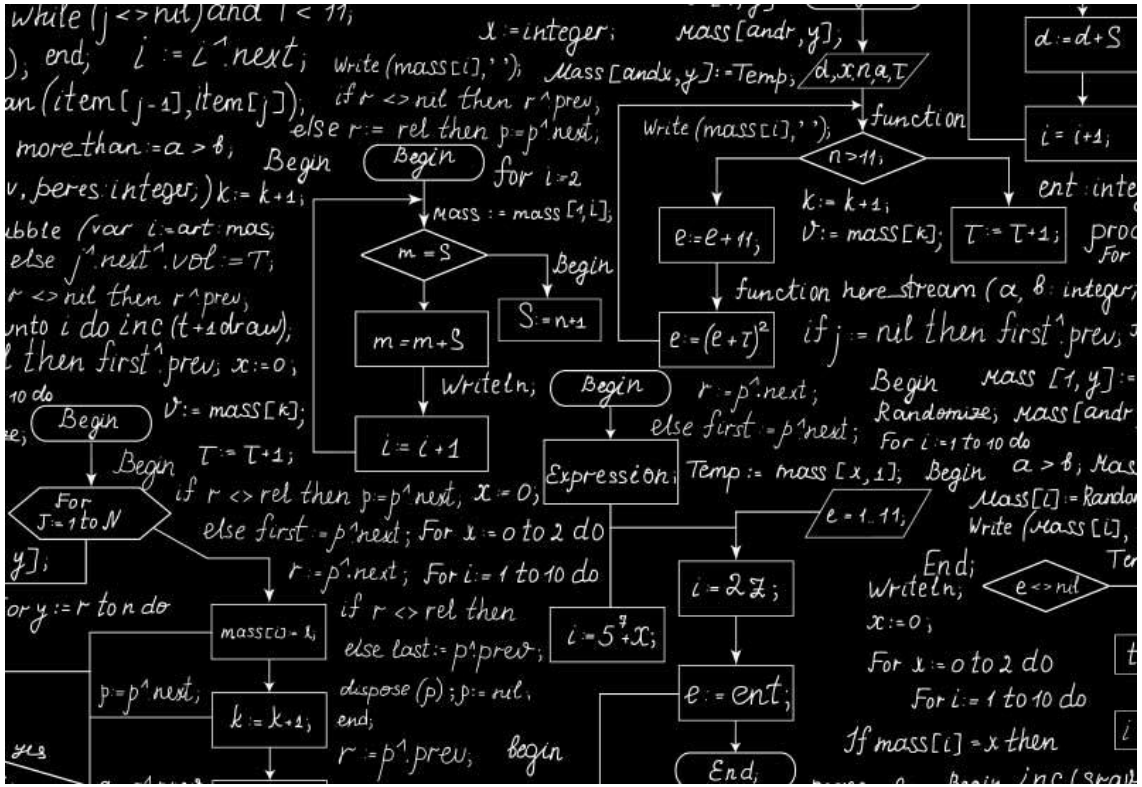
- 4 - בדיקת הנתונים (outliers / group)
  - Data Visualization
  - שיטות סטטיסטיות
  - Projection methods - PCA

Feature Selection

- Feature Selection - בחירת המאפיינים השימושיים לפתרון הבעיה
- שיטות לביצוע Feature Selection

1. Filter Methods - שימוש בשיטות סטטיסטיות לדירוג המאפיינים
2. Wrapper Methods - השוואת הדיוק בין מודלים עם מאפיינים שונים
3. Embedded Methods - סינון מאפיינים ע"י רגולריזציה (LASSO, Elastic Net and Ridge Regression)
   - לכל מודל ניתן לבצע רגולריזציה בשיטות שונות, לדוג' בעצי החלטה ע"י הגבלת עומק העץ

הרצת אלגוריתם ראשוני



- Test Harness
  - שיטות חלוקה ל-train & test (Cross validation, K-fold)
- **בחירת מדד לביצועי המכונה (precision, recall) - בהתאם לבעיה יש להחליט מה חשוב יותר**
- בחירת אלגוריתים מתאים

שיפור המודל

1. שיפור ה-Data
    1. הוספת נתונים
    2. הורדת נתונים
    3. שינוי דרך הצגת המאפיינים
2. השוואה בין אלגוריתמים שונים
3. Algorithm Tuning

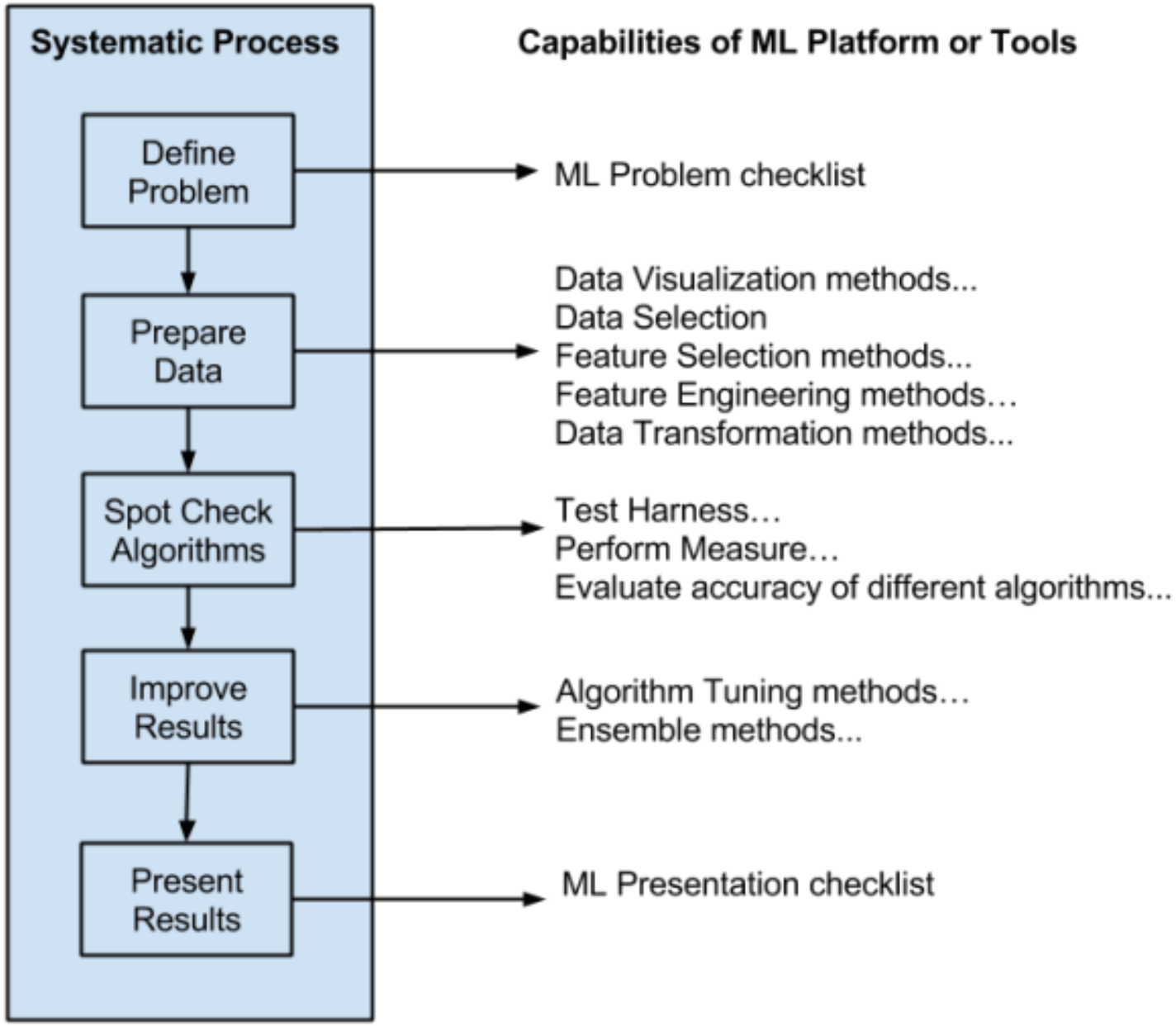
שיפור המודל

4. Ensembles – שילוב של מספר מודלים מוצלחים, לקבלת תוצאה אופטימאלית

- Bagging - אימון מודל זהה על דוגמאות שונות מתוך ה - training set
- Boosting - אימון מודל זהה על דוגמאות שונות מתוך ה-training set בשרשור, ומתן דגש על לימוד דוגמאות שסווגו לא נכון
- Blending – הכנסת תוצרי אימון של מודלים שונים כקלט ולימודם ע"י אלגוריתים חדש לקבלת תוצאה משוקללת
- מומלץ להשתמש בשיטות Ensembles רק לאחר מיצוי שאר השיטות

הצגת התוצאות

- סיכום תוצאות
  - הקשר (למה?) - מה קיים כיום והמוטיבציה למחקר
  - בעיה (שאלה) - תיאור הבעיה ע"י שאלה
  - פתרון (תשובה) - תיאור הפתרון ע"י תשובה לשאלה שנשאלה
  - ממצאים - רשימת הממצאים העיקרית. הממצאים יכולים להיות ב- Data בשיטה או במודל
  - מגבלות המחקר - איזה Data דרוש, מתי המודל לא עובד, רמת האמינות של המודל
  - מסקנות (למה + שאלה + תשובה) - חזרה על העיקרים בצורה תמציתית וניתנת לזכירה


Systematic Process
Capabilities of ML Platform or Tools
Define Problem
ML Problem checklist
Prepare Data
Data Visualization methods...
Data Selection
Feature Selection methods...
Feature Engineering methods…
Data Transformation methods...
Spot Check Algorithms
Test Harness…
Perform Measure…
Evaluate accuracy of different algorithms...
Improve Results
Algorithm Tuning methods…
Ensemble methods...
Present Results
ML Presentation checklist

לקריאה נוספת

- Machine Learning Checklist.pdf
- https://machinelearningmastery.com/process-for-working-/through-machine-learning-problems
- https://machinelearningmastery.com/machine-learning-/checklist