Causal AI models for retail sales prediction

Models

Избрани се 4 множества на податоци, врз кои се извршени 4 типа на каузална анализа, користејќи соодветни модели. Бидејќи станува збор за податоци од типот на продажба, множествата што може да се најдат од овој тип не се случајно избрани. Ова значи дека не можат да се извршат анализи од типот Randomized Control Trials (RCT) – и тоа е земено во предвид кога се извршуваат анализите. Користиме Difference-in-Differences модел, кој што ги споредува податоците после и пред интервенција, при што не ни е потребна рандомизација, Propensity Score Matching модел, кој што го елиминира selection bias и Regression Discontinuity Design (RDD) кој што работи врз податоци кои имаат точка што ги разделува помеѓу интервенирани и не интервенирани. Исто така, поради недостатокот на „интервенции“ во самите дата сета, истите моравме да ги симулираме.

Првото множество на податоци е Warehouse Retail Sales, користејќи модел за DiD (Difference-in-Differences) анализа. Бидејќи за оваа анализа е потребна некаква интервенција, но во множеството на податоци оваа интервенција не постои, ние ќе измислиме интервенција, претпоставувајќи дека истата се случува во Февруари 2019. Како главна група, се избираат сите продажби кои што се од тип на вино, а контролна група се сите останати (пиво и ликер). Променливата што ја мериме се total_sales. Важно е да се напомене дека во множеството на податоци има доста пропуштени редови – пример нема податоци за одредени месеци во 2017 и 2018, што би влијаело врз конечниот резултат, доколку ова беше вистинска интервенција.

После анализата, ги добиваме овие заклучоци:
Коефициенти го репрезентираат ефектот врз променливата total_sales.

- За контролната група, пред интервенцијата. добивме коефициент 58710.
- За главната група тип вино, пред интервенцијата, добивме -23980.
- За двете групи, после интервенцијата, добивме коефициент 1925, со p вредност 0.510

- DiD_interaction коефициент е -6603 со p вредност 0.115(што претставува ефектот на интервенцијата врз главната група споредено со контролната група)
- R square = 0.814 (добар фит на моделот)

Овие резултати ни кажуваат, дека не само што не постои статистички значаен ефект интервенцијата, туку поради негативниот DiD коефициент, интервенцијата всушност ја намалува продажбата.

За второто множество на податоци, го искористив Propensity Score Matching (PSM) моделот, со Scanner Data множеството на податоци. Повторно ќе треба да „измислиме“ сценарио каде што, ќе избереме два типа на SKU_category (X52 и OH2), кои што ќе добијат интервенција (пример намалување на цена), а останатите ќе бидат контролна група. Целта ни е, да видиме, користејќи PSM, да го елемининираме selection bias, правејќи го множеството повеќе како да е случајно избрано. Откако ќе го направиме тоа, ќе можеме да ја тестираме ефикасноста на интервенцијата. Во моделот, прво нешто што правиме откога ќе ги поделиме на две групи, го пресметуваме propensity score (PS), и секој пар од двете групи го поврзуваме соодветно (најсличен PS). Откога ги имаме само поврзаните податоци, можеме да направиме пресметка на средната вредност.

Според конечните резултати добиваме дека:

- Просечната продажба на главната група е 6.58
- Просечната продажба на контролната група е 0.33
- Разликата меѓу нив е 6.25

Може да заклучиме дека измислената промоција значително ги зголеми продажбите за главните SKU групи во споредба со контролната.

Следен модел што ќе го пробаме е causal Directed acyclic graph (DAG) врз третото множество на податоци, Supermart Grocery Sales. DAG е тип на граф што ни овозможува да претставиме каузална врска помеѓу различните променливи во множеството. Претпоставена е најверојатната врска помеѓу

дадените променливи, така што Discount и Region влијаат врз Sales, Sales влијае врз профит, а Category влијае врз Sales и Profit. Овој граф е демонстриран во DAG.ipynb фајлот. Се разбира овие врски се хипотетички, па затоа ќе ги тестираме, користејќи Ordinary Least Squares (OLS) регресија. Споредени се врските меѓу Discounts и Sales, Sales и Profit, и Region и Sales. Соодветно добивме:

- Discounts -> Sales: coeff: -42.66, p-value: 0.582
- Region->Sales: 1.9874, p-value: 0.605
- Sales->Profit: 0.2515, p-value: 0.000

Може да заклучиме дека p-вредностите на Discount и Region со Sales се премногу големи и над 0.05, односно тие немаат значително влијание врз Sales променливата. Додека пак Sales има значително влијание врз Profit. Ова ни кажува дека замисленият DAG не е најдобра претстава за каузалната врска помеѓу променливите, и потребни ни се дополнителни променливи со кои би требало да направиме попрецизен DAG.

За множеството на податоци Retail Sales, извршена е Regression Discontinuity Design (RDD) анализа, користејќи ја цената по единица производ како главна променлива, со пресечна точка на $300. Анализата покажа:

- Јасна разлика помеѓу двете групи кај точката на $300
- Ефектот на интервенцијата е $11.73
- Просечна продажба под $300 – $87.91
- Просечна продажба над $300 – $1017.42
- Разлика од $924.51

Иако големата разлика би сугестирала дека има огромна разлика помеѓу високи и ниски цени на продуктите, со RDD заклучуваме дека ефектот на поставување на цена на продукт над $300 е прилично мал ($11.73). Со други зборови производите над $300 ќе генерираат $11.73 повеќе во вкупна распродажба.