1. Look at the frequencies of each category

Afbeelding met tekst

Automatisch gegenereerde beschrijving

1. Look at the number of observations for each category

Afbeelding met tekst

Automatisch gegenereerde beschrijving

Dag Bettina,

Mijn excuses voor deze zwakke opmaak maar ik heb het wat druk op dit moment en wil u toch nog overzicht bieden van mijn problemen. Hierboven ziet u de frequentie van de observaties per categorie (1ste screenshot) en het aantal observaties per categorie (2de screenshot).

Bij de eerste foto is het belangrijk om op te merken dat er grote verschillen zijn tussen de frequenties per categorie. In een ideale situatie, heeft elke categorie namelijk ongeveer evenveel observaties, wat uitkomt op ongeveer 3,4% per categorie. In ons geval is er echter 1 grote categorie (=majority class) en 28 kleine categorieën (= minority classes). Dit zorgt ervoor dat een machine learning model veel meer moeite heeft om deze kleine categorieën te leren voorspellen. Het heeft namelijk veel meer informatie over de majority class en aangezien deze toch 86% van de tijd het geval is, zal het algoritme geneigd zijn vaker deze classe te voorspellen. Dit probleem heet eigenlijk ‘class imbalance’. Hoe meer categorieën er zijn, hoe moeilijker het voor het algoritme wordt om de verschillende klasses te voorspellen. Dit probleem wordt dan nog eens versterkt door de grote class imbalance aanwezig in de dataset. Dit is dan ook de reden waarom ik het aantal klasses zou willen beperken. Voorts vind ik ook terug in de literatuur dat elke extra klasse zorgt voor een vermindering van de performance.

Bij de tweede foto vindt u het aantal observaties per category. Hier is het belanrijk om op te merken dat enkele categorieën minder dan 30 observaties hebben. Om een model op te stellen, splits ik de data in een train en een test set. Hiervan wordt 80% aan de training set toegewezen. In de training set pas ik dan nog eens cross validatie toe, waardoor weer maar 80% van de training set gebruikt wordt door het algoritme om de verschillende klasses te leren. Dit betekent dat er van die 30 observaties of minder maar 24 overblijven na de eerste split, waarvan er slechts 20 gebruikt worden om het model te tunen en 4 om de evaluatie van verschillende parameters te testen. Dit zorgt ervoor dat het algoritme te weinig informatie heeft over deze klasses om accurate voorspellingen te doen hierover en 2 van de 4 test observaties verkeerd geclassificeerd worden heeft dit een grote impcact op de performance.

Dit zijn dus mijn 2 grote problemen waarmee ik ook naar professor Van den Poel zal gaan. Op dit moment zie ik dan ook 2 opties die ik met hem zal bespreken. Naast 1 van deze 2 opties kan ik natuurlijk ook het model met 29 klasses maken en de performance vergelijken. Dit zijn volgens mij de opties:

1. Mijn eerste voorstel is om terug te gaan naar de abstractere levels zoals product/Service, market, marketing, firm scope… Dit zou namelijk het probleem van zowel het aantal categorieën als het aantal observaties aanpakken. Voorts zorgt dit ook nog steeds voor 11 klassen, wat nog steeds als een groot aantal gezien wordt voor classificatiemodellen. Daarnaast heeft elke categorie wel minstens 50 observaties. Dit is dan ook de optie die ik prefereer maar dit is natuurlijk onder voorbehoud van de feedback van professor Van den Poel.
2. Ik kan de categorieën met minder dan 30 observaties herclassificeren naar de categorie ‘None’. Dit zorgt ervoor dat het probleem over het aantal observaties toch wat gereduceerd wordt. Ook dit aantal zou ik natuurlijk met Van den Poel moeten bespreken en het zou kunnen dat dit getal wordt verhoogd naar bv 70 of 100. Persoonlijk denk ik echter wel dat het niet meer dan 70 zal zijn en dat hij het aantal categorieën als een groter probleem zal zien vanaf dat er meer dan 70 observaties zijn. In deze scenario’s zouden dus de volgende categorieën verdwijnen:
3. Tot en met 30:

A1. Product and service deletion

A2. New product or service market entry

A3. Customer service relations

A4. Spin-off

A5. Supplier and outsourcing

1. Extra tot en met 70

B1. Reorganization and restructuring

B2. Licensing

B3. Reducing market presence or exiting a market (product/service/geographical)

1. Tot en met 100:

C1. New geographical market entry

C2. Pricing actions

C3. Joint venture

C4. De-venturing

**Input Prof. Peeters & Bettina:**

Note:

‘None’ does not need to be kept but I told them that this variable can’t be dropped as this is part of our data. Was this correct?

To keep, ranked in order of importance:

* Merger & Acquisition: 575
* Strategic alliance: 996
* External venturing: 224
* Join-venture (if too small, you can add it to the “external venturing” category, but we prefer to keep it separate): 89
* New product/service introduction: 648
* Product/service improvement: 318
* New geographical market entry + new product/service market entry (together if it helps. We like new geographical market entry the most): 84 + 11
* Expanding in existing markets: 311

Ranked in order of importance. Keep, but if needed you can drop:

* Financing: 328
* R&D: 172
* Human resources: 201
* Production: 265
* Corporate governance: 746
* Divestiture: 249
* All marketing categories together (advertising, pricing and customer service): 152 + 77 + 26

To drop:

* De-venturing
* Reorganization and restructuring
* Spin-off
* Product or service deletion.
* Reducing market presence or exiting a market.
* Licensing
* Supplier and outsourcing
* Legal
* Signaling
* symbol