Aalto-yliopisto
Perustieteiden
korkeakoulu

Startup yritysten onnistumisen ennustaminen (valmiin työn esittely)

Tomi Lahti

29.08.2024

Ohjaaja: Jukka Kohonen

Valvoja: Jukka Kohonen

Aalto-yliopisto
Perustieteiden
korkeakoulu

Tausta

- Startup-yritysten menestyksen ennustaminen on tärkeä tutkimuskohde nopeasti kehittyvässä liiketoimintaympäristössä.
- Onnistumisen ennustaminen auttaa yrittäjiä, sijoittajia ja päätöksentekijöitä tekemään informoituja päätöksiä
- Tutkimuksessa tarkastellaan maantieteellisen sijainnin, toimialan ja rahoituksen vaikutusta startup-yritysten menestykseen

Tavoitteet

- Rakennetaan logistinen regressiomalli startup-yritysten menestyksen arvioimiseksi
- Selvitetään, miten startup-yritysten menestystodennäköisyys riippuu rahoituksen, sijainnin ja toimialan kaltaisista keskeisistä tekijöistä.
- Startupin menestyminen kyseisessä datassa tarkoittaa; 0 = yritys on suljettu, 1 = yritys on tehnyt ns. exitin eli yritys on myyty
- Arvioidaan ennustemallien tarkkuutta ja yleistettävyyttä käyttämällä ristiinvalidointia startup-yritysten menestyksen ennustamisessa

Aineisto

- Tutkimuksessa käytetään "Startup data" -aineistoa, joka on peräisin tunnetusta Kaggle-alustalta, joka on datatieteilijöiden ja koneoppimisen ammattilaisten verkkoyhteisö.

- Tässä tutkimuksessa käytetty aineisto sisältää tietoja yhdysvaltalaisista startup-yrityksistä ja niiden ominaisuuksista. (mm. maantieteellinen sijainti, toimiala, rahoitus)

- Aineiston luotettavuuden ja laadun varmistamiseksi on tärkeää huomata, että Kaggle on osa Googlen alaisuutta. Tämä yhteys voi tukea aineiston luotettavuutta ja vahvistaa sen käyttökelpoisuutta.

Aineisto

- Aineisto sisältää 923 riviä ja 43 eri muuttujaa. Suurin osa muuttujista on kaksijakoisia eli binäärimuuttujia, jotka voidaan tulkita myös 0 = False ja 1 = True.

- Aineistossa menestyneiden startupien määrä on suurempi kuin epäonnistuneiden, mikä on ristiriidassa yleisen käsityksen kanssa startup-yritysten menestyksestä. Tämä korostaa datan kriittisen tarkastelun tärkeyttä.

Menetelmä

- **Logistinen regressio**
  - Logistinen regressio on lineaarisen regression jatke, joka mallintaa ennustajamuuttujien ja tietyn lopputuloksen todennäköisyyden välistä riippuvuutta.
  - Selittävinä muuttujina maantieteellinen sijainti, toimiala ja rahoitusmuodot, vastemuuttujana yrityksen menestys (0 = epäonnistunut, 1 = menestynyt)

$$P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n)}}. \qquad (1)$$

- Malli tuottaa estimaatit, keskivirheet ja Z-arvot, joiden avulla arvioidaan ennustajamuuttujien tilastollista merkittävyyttä.
- Kertoimet voidaan tulkita siten, että jos kerroin on suurempi kuin 1, muuttuja lisää onnistumisen todennäköisyyttä, ja jos se on pienempi kuin 1, se vähentää sitä, ja suuret absoluuttiset Z-arvot (yli 1,96 tai alle -1,96) viittaavat, että muuttuja on tilastollisesti merkitsevä.

Menetelmä

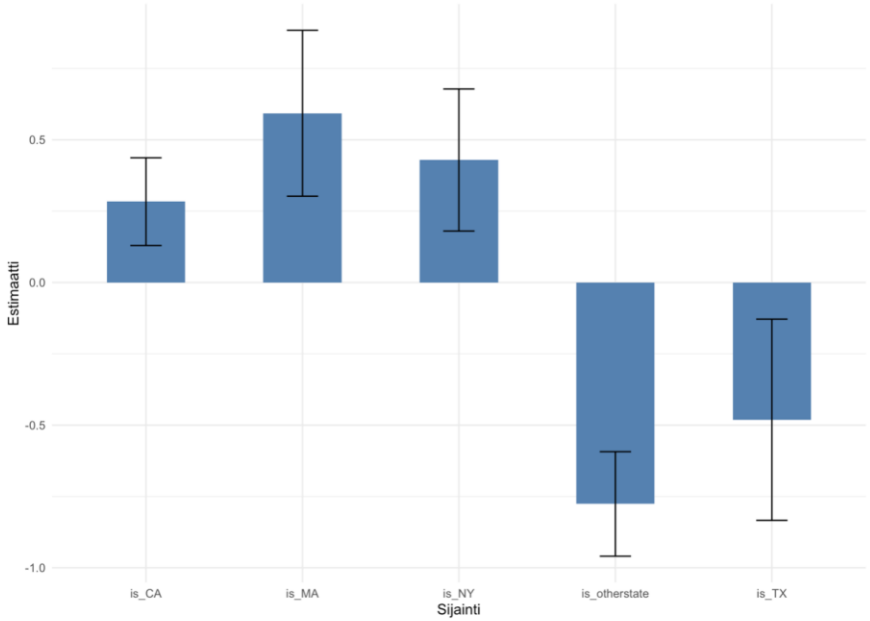
- **Logaritminen pisteytys**
  - Logaritminen pisteytys on pisteytysääntö, joka kvantifioi todennäköisyysennusteiden oikeellisuuden. Se on erityisen hyödyllinen binääritapahtumissa, joissa ennustetaan kyllä- tai-ei-tuloksia
  - Kaava logaritmisen pisteytyksen laskemiselle kaikille havainnoille:

$$\text{Log score} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \cdot \log(p_i) + (1 - y_i) \cdot \log(1 - p_i)\right]. \qquad (2)$$

- On tärkeää huomata, että koska logaritmi negatiivisella todennäköisyydellä (joka on välillä 0 ja 1) on negatiivinen, kaavassa oleva miinusmerkki kääntää nämä arvot positiivisiksi, mikä tarkoittaa, että pienempi logaritminen pistemäärä on parempi.

Tulokset

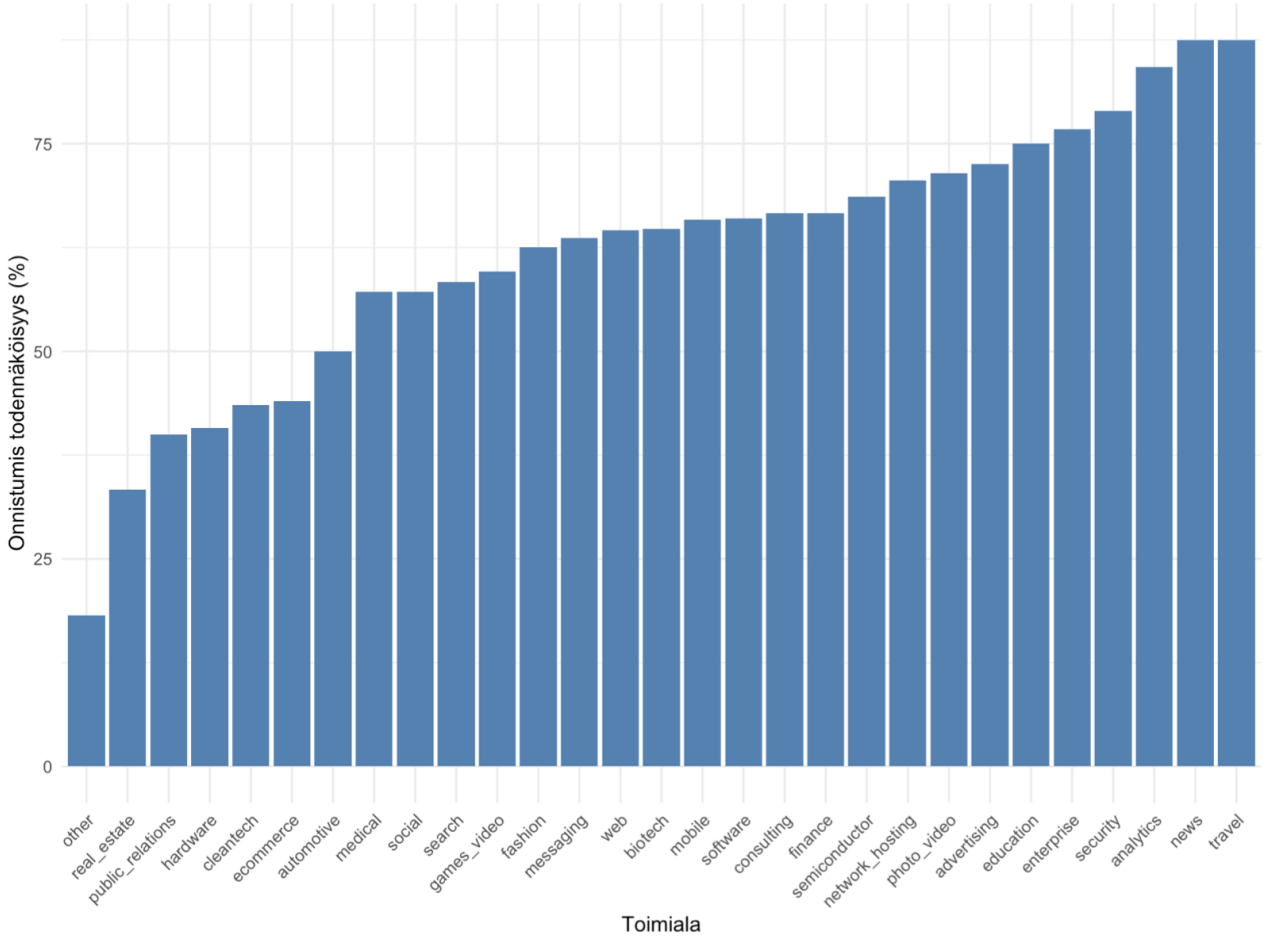
Maantieteellinen sijainti



	Estimaatti	Keskivirhe	Z-arvo	p-arvo
(Intercept)	0,4293	0,1094	3,925	0,00009
Kalifornia	0,2832	0,1536	1,843	0,06533
New York	0,4290	0,2490	1,7233	0,08483
Massachusetts	0,5929	0,2906	2,0401	0,04134
Texas	-0,4810	0,3524	-1,3647	0,17235
Muut osavaltiot	-0,7759	0,1831	-4,2383	0,00002

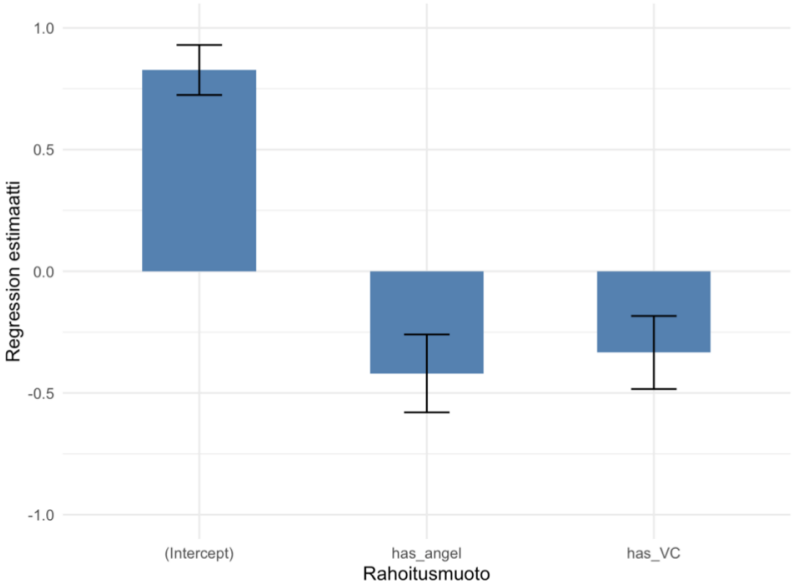
Tulokset

Toimiala



Tulokset

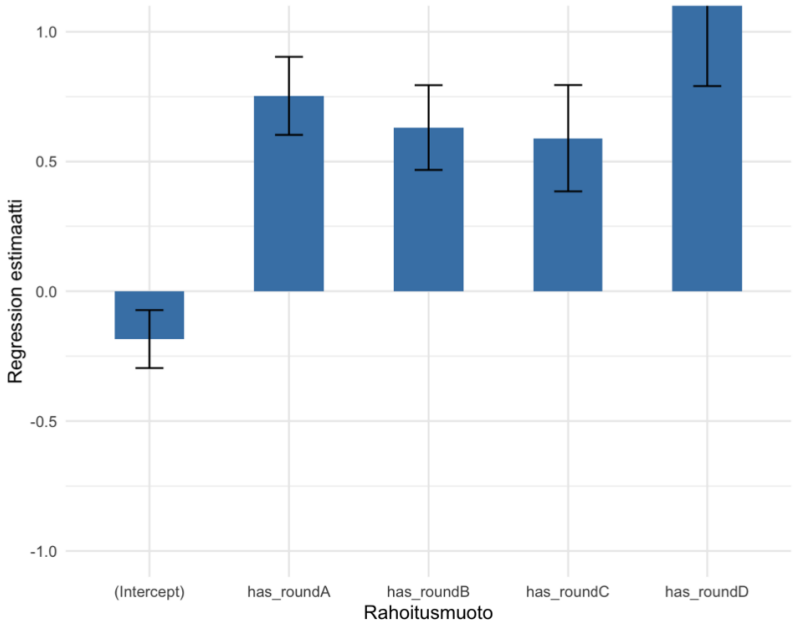
Rahoitus – riskipääoma sekä enkelit rahoitusmuotona



Muuttuja	Estimaatti	Keskivirhe	Z-arvo	P-arvo
Intercept	0,8272	0,102	8.146	$< 0{,}001$
has_VC	-0,3338	0,150	-2.229	0,026
has_angel	-0,4196	0,160	-2.628	0,009

Tulokset

Rahoitus – rahoituskierrosten määrän vaikutus



Muuttuja	Regression estimaatti	Keskivirhe	Z-arvo	P-arvo
Intercept	-0,1843	0,1114	-1,6498	$< 0{,}001$
has_roundA	0,7532	0,1503	5,0368	0,0000
has_roundB	0,6305	0,1637	3,8838	0,0001
has_roundC	0,5895	0,2052	2,9223	0,0035
has_roundD	1,1257	0,3347	3,5546	0,0004

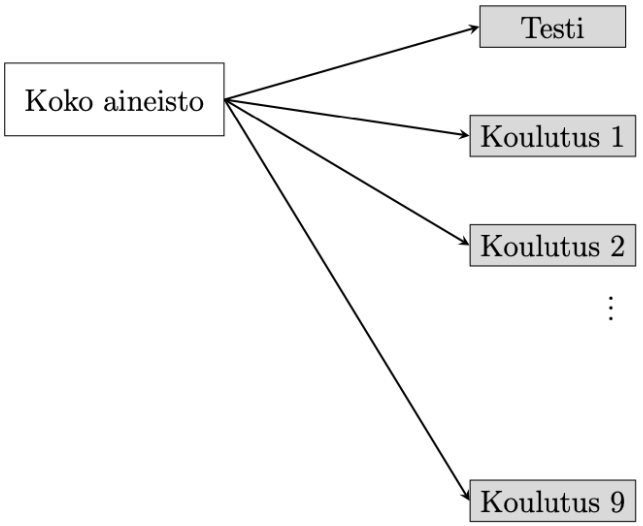
Tulokset

Tulosten yhteenveto

- Massachusettsin (MA) startupit menestyvät tilastollisesti merkittävästi paremmin; Texas (TX) ja muut osavaltiot heikommin.
- Uutis- ja matkailualan startupit osoittavat korkeimmat onnistumisprosentit (87,5 %), mutta otoskoko on pieni; analytiikka ja tietoturva myös menestyviä, mutta vähemmän yrityksiä.
- Riskipääoma ja enkelisijoitukset liittyivät yllättäen negatiivisesti menestykseen, mahdollisesti tiukkojen ehtojen vuoksi.
- A-, B-, C- ja erityisesti D-kierrokset lisäävät merkittävästi startupien menestymisen todennäköisyyttä.

Ennustuskyvyn arviointi

- Ristiinvalidointi on käytetty menetelmä mallin ennustuskyvyn arviointiin, jossa data jaetaar kymmeneen osajoukkoon (folds).
- Yhdeksää osajoukkoa käytetään mallin koulutukseen ja yhtä testaukseen, prosessi toistetaan kymmenen kertaa niin, että jokain osajoukko toimii kerran testidatana.
- Mittaa mallin yleistä ennustuskykyä ja auttaa tunnistamaan mahdolliset ylisovittamisen ongelmat sekä mallin suorituskyvyn vaihtelut osajoukoissa.



Ennustuskyvyn arviointi

Ristiinvalidointi kolmelle mallille

Malli	Tarkkuus (%)	Log-pisteytys
A (Sijainti)	63,92	0,6374
B (VC ja enkelit)	64,68	0,6467
C (Rahoituskierrokset)	67,71	0,6043

A: Tarkkuus 63,92 %, logaritminen pisteytys 0,6374. Ristiinvalidointi paransi hieman logaritmista pisteytystä, mutta vaikutus oli vähäinen.

B: Ristiinvalidoinnin jälkeen logaritminen pisteytys nousi hieman 0,6467:ään, mikä viittaa mallin parantuneeseen yleistettävyyteen.

C: Paras tarkkuus 67,71 % ja pienin logaritminen pisteytys 0,6043. Ristiinvalidointi vahvisti mallin ennustuskykyä, mikä osoittaa rahoituskierrosten merkittävyyden ennustamisessa.

Yhteenveto

- Startupien menestyksen ennustaminen on monimutkaista ja vaatii useiden tekijöiden ymmärtämistä.
- Tietyt osavaltiot, kuten Massachusetts, näyttivät merkittäviä vaikutuksia, mutta sijainnin vaikutus menestykseen on monitahoinen.
- Tietyt alat, kuten uutis- ja matkailuala, osoittivat korkeaa menestystä, mutta tulokset perustuvat rajalliseen otokseen.
- Perinteiset rahoituslähteet eivät osoittaneet selkeää positiivista vaikutusta, mutta rahoituskierrosten määrä korreloi vahvasti menestyksen kanssa.
- Ristiinvalidoinnin malli C oli vahvin ennustaja, mikä korostaa jatkuvan taloudellisen tuen merkitystä.

Viitteet

- Z. J. Acs, S. Estrin, T. Mickiewicz, and Ll. Szerb. Institutions, Entrepreneurship and Growth: The Role of National Entrepreneurial Ecosystems. *SSRN Electronic Journal*, 2014.
- M. Anderson. MIT: Median Error Measurement in Large News Outlets Using Machine Learning - Unite.Ai. n.d.
- Manish Kumar Chauhan. Startup success prediction dataset, 2024.
- B. Chernev. What Percentage of Startups Fail?, 2020.
- S. Enginsoy. Analysis of the North America Startup Ecosystems, 2023.
- E. L. Glaeser, W. R. Kerr, and G. A. M. Ponzetto. Clusters of Entrepreneurship. *Journal of Urban Economics*, 67(1):150–168, 2010.
- T. Gneiting and M. Katzfuss. Probabilistic Forecasting. *Annual Review of Statistics and Its Application*, 1:125–151, 2014.
- P. Gompers and J. Lerner. The Venture Capital Revolution. *Journal of Economic Perspectives*, 15(2):145–168, 2001.
- T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning - Data Mining, Inference, and Prediction, Second Edition.* Springer Series in Statistics, 2017.
- Katarina Juvonen. Leadership Do's and Don'ts For Fast Growth. *Finnish Startup Community*, 2023.
- MakeUseOf. Beginner's Guide to Kaggle, 2023.
- C. Mason and R. Brown. Creating Good Public Policy to Support High-Growth Firms. *Small Business Economics*, 40(2):211–225, 2011.