Spelling errors in peer-to-peer (P2P) loan applications and probability of default

Michelle Seng Ah Lee (Supervisor: Jat Singh)

Outcomes

An analysis of P2P lending data revealed statistically significant associations between spelling errors and non-standard language (slang and abbreviation) and a higher default risk in peer-to-peer lending.

For example:

P(default | 0 spelling error) = 17.4%

P(default | 1 spelling error) = 19.9%

The text length, type of error (orthographic or phonological), and the distance of the typo from the correction do not affect the probability of default.

Next steps and ethical / legal implications

Literature on causes of spelling errors identifies three potential drivers:

Dyslexia

Non-native speaker

Carelessness

Intelligence and education level are not reliable predictors of poor spelling. Credit discrimination on the basis of disability status and national origin is prohibited in the U.S. (ECOA 1974).

Future research will address this methodological challenge: How do we isolate the impact of carelessness?

Methodology

Data: Lending Club (2007-2011)

42,309 loan applications and their descriptions, 15% default rate

Below is a partial excerpt from one of the descriptions, with non-standard language in *italics* and accidental typos in **bold**:

> Borrower added on 07/21/11 paying off house and 4,000.00 back to my saving *acct* bought a 12,000.00 boat and **payed** cash for it. Borrower added on 07/24/11 reason for paying house off i never have more then one loan at a time. Borrower added on 08/02/11 Thanks to the investors you're making a sure thing with me. my credit rating has **allways** been number one with me. thanks again

Features engineered:

Facilitated by open source packages (e.g. soundex, enchant) but required extensive manual review and correction

- Spelling error candidates (7,997 unique) identified, false positives (299), medical terms (36), named entities (1,891), acronyms (645) flagged
- Non-standard language (664 slangs, 76 abbreviations) identified
- Phonetic equivalence calculated
- Levenshtein distance calculated

Algorithm: multivariate logistic regression to predict default controlling for income and loan amount

Contact

Michelle Seng Ah Lee

PhD Candidate

University of Cambridge

sal87@cam.ac.uk