University of Ljubljana – Faculty of Economics
Subject: Business Intelligence
Mentor: **prof. dr. Jurij Jaklič**

Study Program: Business Informatics (Masters)
School Year: 2016/2017

Seminar Paper

as part of subject

BUSINESS INTELLIGENCE

Personality type prediction with text mining

Student: Sara Jakša
Enrollment number: 19524001

Ljubljana, January 2016

Contents

List of Tables

Introduction

1 Justification

1.1 Business Problem

It can be really expensive to replace the employee. Torres (2015) in her article in the Harvard Business Review gives an estimate of $12,500. But some other places give an estimation of up to 200% of the workers yearly salary for the most specialized and highest positions (Borysenko, 2015; Kantor, 2016).

I had a situation like that. I was accepted for the student job. The month later, I was let go, and the reason they gave me was that I was not extroverted enough for the job functions that I would have in the future, even though I was preforming above expectations.

If they would ask me about it, I could have told them this in the interview. But I could do this because first, I know quite a bit because of the different models of the personality types and second, because I did not want a job for any price, even trying to obscure the truth.

But there are cases, where people are not familiar with the concepts (khimhoe, 2012) or they might even have the wrong idea about them. Also some people are desperate, and would be willing to bend the truth in order to get the job.

62% of HR people use personality tests to help them decide about the new hires. 80% of new hires at Fortune 500 are given a a personality test. (Chatterjee, 2015). The rise in the number of companies employing them is because they are lower turnover (Weber, 2015).

But these tests can cause money. Some tests can also be administrated by psychologists. The others require payment for each test that was administrated. Some of the smaller companies can not afford to do this, and it can even save money for the bigger company.

They are also useful, because it is more expensive to hire a bad hire than to not hire a good one. This test could be used as one of the beginning filters for the new job applicants and with that make a process of decision making more efficient. Or it could be used as an additional piece of information, which will allow for the better hires.

1.2 Cost-Benefit Analysis

Table 1: *Cost-Benefit Analysis*

Cost/Benefit	Description	Size
Benefit	Better Hires	$12,500*0.31*number of hires
Benefit	Personality test cost	number of hires*$50
Cost	Cost of Development	$68,000
Cost	Cost of Maintenance	$68,000

For better hires, I have used the $12,500 cost, that the company is not going to have for each bad hire they don't hire (Torres, 2015). Personality can on average predict around 31% of variance of performance (Batty, 2013). There is also a cultural fit, but this is harder to calculate and depends on the company. So this application could increase the accuracy of hiring decision for up to 31%. Also this cost is averted every time the bad hire is avoided, so it gets multiplied by the number of hires.

The personality tests can cost from $50 upwards (CPP Inc., 2017). This number gets multiplied by the number of hires, as it is averted each time the test does not get administrated.

Cost of development was calculated indirectly. This application will probably be the size of a normal app for phone, but used on the computer. The normal cost of an iApp is $200,000, based on the $150 per hour cost of developer(StackOverflow, 2010). But the average cost of software developer is $26.14 (PayScale, 2017). Around half of this was work, so by adjusting the prices, we get the cost of $18,000. Also, since this is not an online app with multiple users, the cost of infrastructure is cut in half (SavvyApps, 2015). This makes the cost of infrastructure $50,000. So the full cost is $68,000.

Cost of maintenance over time is usually the same as the cost of development.

Assuming these numbers, the break even point would be reached, if there are more than 36 people employed in the time we are using this system. Which is quite a huge number for a small company, but minimal for a medium or big company.

This cost benefit analysis assumes that there are people working on it as part of their job and it is going to be used by multiple people.

These numbers look very different, when this is a project that I just want to learn something from. Which in this case is true.

2 Planing

2.1 Data Sources

There are at least three different data sources, that I know of, that they can be used for this project. This are

- Personality cafe forum
- MBTI and MBTI types subreddits
- Blogs of people on Tumblr that have the MBTI type written in their description

I originally wanted to use data from Tumblr, but I ended up using data from Personality cafe, because it would take my more than a month to get the same amount of data from Tumblir, than I was able to get from Personality cafe in 48 hours.

2.1.1 Personality Cafe

The forum Personality Cafe (http://personalitycafe.com/) is a forum, where the personality types enthusiasts gather and discuss different things. The forum allows for the users to define their own personality types. Then this type is displayed under the user name on the subformus that relate specifically the the personality type.

They have no API or something similar, but it is relatively easy to scrape the forum manually and get the data desired.

2.1.2 MBTI Subreddits

Reddit has a couple of communities connected to the MBTI type. There is at least one general (https://www.reddit.com/r/mbti/) and one for each subtype (for example: https://www.reddit.com/r/INTP/). In these subreddits, quite a lot of users have their type displayed next to their name. I did not notice this on the other subreddits. Since these users also post on other subreddits, there is a way to get various posts from them.

Reddit has an API, that could be used for this.

2.1.3 Tumblr

Tumblr (http://tumblr.com/) is a site where people post things. An interesting thing that I noticed is, that a lot of times, people put their MBTI type in their description. There are hundreds of users for each type and they usually have multiple posting.

Tumblir also has an API that can be used for this.

2.1.4 Other possible sources

Different MBTI types also use mailing lists to stay in touch (for example http://lists.vt11.net/mailman/listinfo/intp). Most of these, that I checked out have a archives that are available by simply subscribing to them.

Also, I heard that they are Facebook pages that are devoted to these, but I have not checked these ones out.

2.2 Data Retrieved

After getting data from the sources, it included cleaning.

I removed everything, so all I ended up with was the text of the post and the MBTI type of the person that post it.

Since there was a huge inequality of the classes, I decided to use the number of elements from each class equal to the number of elements in the class with lowest number of element. Since the biggest class had 100x more element, I figured that there was a need for some adjusting.

3 Analysis

3.1 Tools

I realized that I had at least two ways to analyze data. One of them was by using Orange (http://orange.biolab.si/), a data analyzing software. The other was by using the sci-kit and numpy libraries in Python. I first tired to use the Orange, but since I was constantly getting the error messages, I decided move to sci-kit.

3.2 Features to Analyze

Here I had a different ideas of what I could analyze. One of the was the word frequency analysis, which was shown in every book and almost all tutorials. Then I could analyze the use of positive and negative words, the use of punctuation and upper case words, the length of the text, sentence and words, the lexical diversity and many other.

I decide that for this project I am going to be using word frequency analysis, because this is something that I have not done before.

3.3 Models to Use

I have decided to try different models. I started with techniques for Linear CSV, Decision Tree, Naive Bayes, Logistic regression and K-nearest neighbor. I removed the K-nearest neighbor, because I did not have enough RAM to calculate it on my data set.

For each of these, I have tired a couple of different models in order to get the best result on the test data, which was the 30% of data, that I randomly put aside each time for testing.

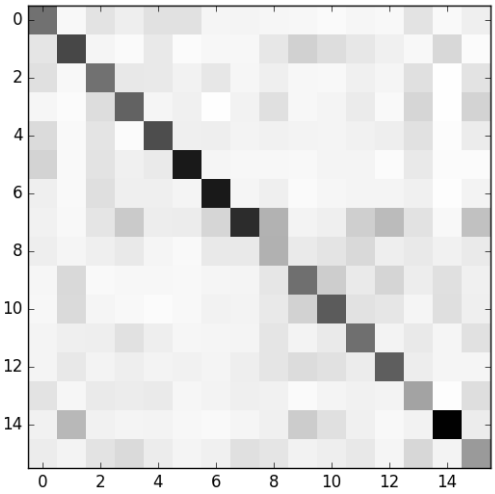
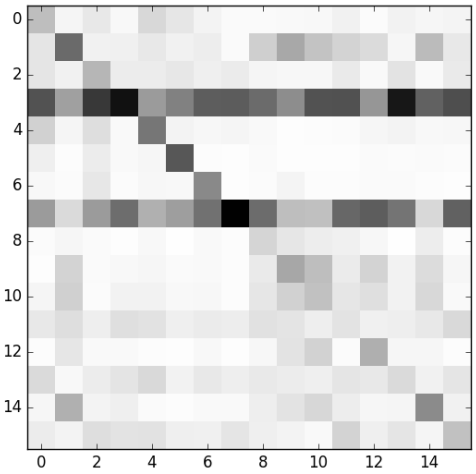
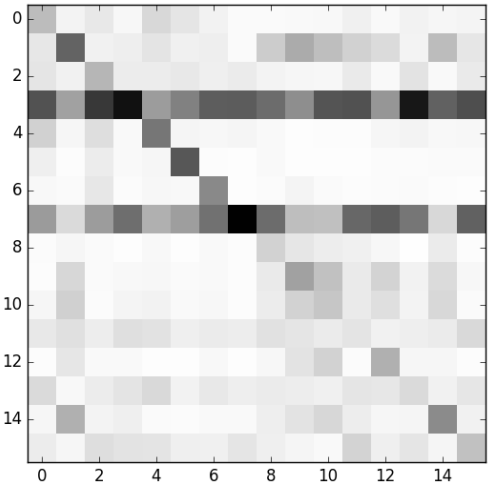
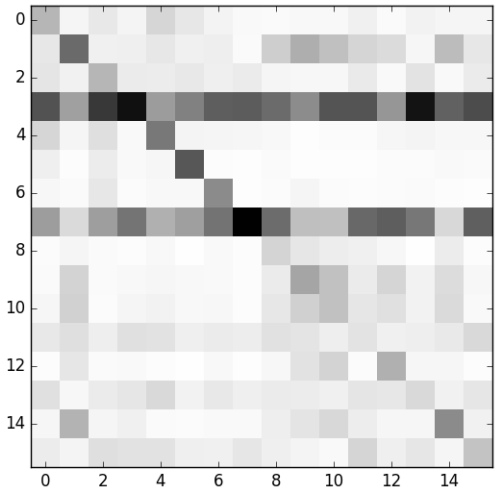
I then decided to try and pit all of them against each other. The following table presents the ration of the texts, that the model was able to classify correctly. Since there are 16 possible classes, the model would need to be higher than 6,25%.

Table 2: *Performance of Model*

Model	Correct Classifications
LinearSVC	33,86%
Decision Tree	19,15%
Naive Bayes	19,26%
Logistic Regression	19,26%

All models are above the random chance of guessing. I tried to combine some of them, but I did not get any better results.

I also checked if there is any bias that these models are making. You can see the pictures below. LinearSVC was the one, where there was no noticeable one, so I decided to use it going forward.

3.4 Trying to Use Personality Type Theory

I then used the theory to try and predict different type information from the data and then combining the models. But I got worse results, so I ended up not using it.

3.5 Interpreting the Model

I then tried to extract the features to try and figure out which words are the best at predicting different types.

Table 3: *Interpretation of the Model*

Type	Use it more than average	Use it less than average
ENFJ	kintsugi, enfj, thorweeps, isotropic, vigro	mbti, istj, applies, istj, sparkles
ENFP	enfp, deployed, snowbell, rebound,pokerface	confidence, boomonster, pinkrasputin, video, afraid
ENTJ	entj, i_really_hate_decisions, nihilo, kuromi, moderator	pretty, infj, enfj, afraid, place
ENTP	thoughtfulthinker, lord, snark, grandmaster, singularity	guys, intj, lots, convince, situation
ESFJ	esfj, islandlight, zdd, gossip, quest	niss, rave, scared, value, se
ESFP	esfp, digger, bozo, dora, cooking	fe, plans, nt, curiosity, purse
ESTJ	estj, thisisntfaith, gal, tzara, metalme	younger, doubt, frens, times, esfj
ESTP	demond, in2itive, grooveshark, ablysmal, dc,	currently, life, hey, cool, supposed
INFJ	infj, yoshi, smooth, giggling, tempting	esfp, infp, esfj, estj, intp
INFP	feelery, daisychain, unicorn, tuttle, pink	esfj, dinner, shit, current, guy
INTJ	rhetorical, kick, chaos, cleo, robin	favorite, istp, moments, trying, esfj
INTP	xenos, systems, switch, boring, pie	sir, isfj, hope, posted, job
ISFJ	rant, ham, nap, cleo, program	conversation, type, thanks, asking, originally
ISFP	loaf, devil, polish, niccolo, faces	sela, home, theory, feelers, enthusiast
ISTJ	pinkrasputin, ama, boomonster, sparkles, memphisto	clearly, prayers, moves, deeply, lol
ISTP	suspected, perceiving, confess, researching, spoke	share, entj, rave, husband, personally

There are more than these words, but these ones have the highest weight for classifying in each type.

But this information is not useful like that, expect to some researchers. It is a lot more useful to have an app to be able to calculate this on the spot.

I have written a small prototype app in python and qt, where a person can simply put the text in and get the most probably type out. It can be found on https://github.com/sarajaksa/schoolwork/tree/master/personalitytype.

References

Batty, R. (2013). Intelligence matters more than you think for career success. Retrieved January 15, 2017, from https://80000hours.org/2013/05/intelligence-matters-more-than-you-think-for-career-success/

Borysenko, K. (2015). What was management thinking? the high cost of employee turnover. Retrieved January 3, 2017, from https://www.eremedia.com/tlnt/what-was-leadership-thinking-the-shockingly-high-cost-of-employee-turnover/

Chatterjee, C. (2015). 5 personality tests hiring managers are using that could make or break your next job interview. Retrieved January 3, 2017, from http://www.msn.com/en-nz/money/careersandeducation/5-personality-tests-hiring-managers-are-using-that-could-make-or-break-your-next-job-interview/ar-BBl1TRB

CPP Inc. (2017). The myers-briggs type indicator assesment. Retrieved January 15, 2017, from https://www.mbtionline.com/TaketheMBTI

Kantor, J. (2016). High turnover costs way more than you think. Retrieved January 3, 2017, from http://www.huffingtonpost.com/julie-kantor/high-turnover-costs-way-more-than-you-think_b_9197238.html

khimhoe. (2012). Introvert versus extrovert. Retrieved January 3, 2017, from http://www.plinky.com/answers/180870

PayScale. (2017). Hourly rate for industry: software development. Retrieved January 15, 2017, from http://www.payscale.com/research/US/Industry=Software_Development/Hourly_Rate

SavvyApps. (2015). How much does an app cost: a massive review of pricing and other budget considerations. Retrieved January 15, 2017, from https://savvyapps.com/blog/how-much-does-app-cost-massive-review-pricing-budget-considerations

StackOverflow. (2010). How much does it cost to develop an iphone application? [closed]. Retrieved January 15, 2017, from http://stackoverflow.com/questions/209170/how-much-does-it-cost-to-develop-an-iphone-application/3926493

Torres, N. (2015). It's better to avoid a toxic employee than hire a superstar. Retrieved January 3, 2017, from https://hbr.org/2015/12/its-better-to-avoid-a-toxic-employee-than-hire-a-superstar

Weber, L. (2015). Today's personality tests raise the bar for job seekers. Retrieved January 3, 2017, from http://www.wsj.com/articles/a-personality-test-could-stand-in-the-way-of-your-next-job-1429065001