

**PRAGUE UNIVERSITY OF
ECONOMICS AND BUSINESS**
FACULTY OF ACCOUNTING AND FINANCE
Department of Banking and Insurance



**Application of Machine Learning
Algorithms within Credit Risk
Modelling**

Master's thesis

Author: Bc. Petr Nguyen

Study program: Banking and Insurance | Data Engineering

Supervisor: prof. PhDr. Petr Teplý, Ph.D.

Year of defense: 2023

Declaration of Authorship

I, as an author, hereby declare that I wrote and compiled the Master's thesis
"Application of Machine Learning Algorithms within Credit Risk Modelling"
independently, using only the resources and literature listed in bibliography.

May XX, 2023, Prague

Petr Nguyen

Abstract

The abstract should concisely summarize the contents of a thesis. Since potential readers should be able to make their decision on the personal relevance based on the abstract, the abstract should clearly tell the reader what information he can expect to find in the thesis. The most essential issue is the problem statement and the actual contribution of described work. The authors should always keep in mind that the abstract is the most frequently read part of a thesis. It should contain at least 70 and at most 120 words (200 when you are writing a thesis). Do not cite anyone in the abstract.

Keywords: machine learning, data science, credit risk, probability of default, loans, mortgages

Abstrakt

Nutnou součástí práce je anotace, která shrnuje význam práce a výsledky v ní dosažené. Anotace práce by neměla být delší než 200 slov a píše se v jazyce práce (tj. česky, slovensky či anglicky) a v překladu (tj. u anglicky psané práce česky či slovensky, u česky či slovensky psané práce anglicky). Anotace práce by neměla být delší než 200 slov a píše se v jazyce práce (tj. česky, slovensky či anglicky) a v překladu (tj. u anglicky psané práce česky či slovensky, u česky či slovensky psané práce anglicky). V abstraktu by se nemělo citovat.

Klíčová slova: machine learning, data science, kreditní riziko, pravděpodobnost defaultu, úvěry, hypotéky

Acknowledgments

I, as an author, would like to express my deepest gratitude and thanks to my supervisor prof. PhDr. Petr Teplý, Ph.D. for his help and significant advice throughout my thesis. Last but not least, I would like to also thank to my family for an enormous support during my studies.

Contents

List of Tables	viii
List of Figures	ix
Acronyms	x
1 Introduction	1
2 Credit Risk Modelling	2
2.1 Formal requirements of master's thesis	2
2.2 Template adjustments and meta-data	3
3 Machine Learning	4
3.1 Terminology	5
3.2 Algorithms	5
3.2.1 Logistic Regression	5
3.2.2 Decision Tree	5
3.2.3 Logistic Regression	5
3.2.4 Naive Bayes	5
3.2.5 K-Nearest Neighbors	5
3.2.6 Random Forest	5
3.2.7 Gradient Boosting	5
3.2.8 Support Vector Machine	5
3.2.9 Neural Networks	5
3.3 Evaluation Metrics	5
3.3.1 Confusion Matrix	5
3.3.2 Accuracy	5
3.3.3 Recall	5
3.3.4 Precision	5
3.3.5 F1 Score	5

3.3.6	AUC	6
3.3.7	Kolmogorov-Smirnov	6
3.3.8	Somer's D	6
3.3.9	Matthews Correlation Coefficient	6
3.3.10	Brier Score Loss	6
3.3.11	Jaccard Score	6
3.3.12	Zero-One Loss	6
3.4	Hyperparameter Tuning	6
3.4.1	Grid Search	7
3.4.2	Random Search	7
3.4.3	Bayesian Optimization	7
3.5	Imbalanced Class Distribution	7
3.5.1	Random Oversampling	7
3.5.2	SMOTE Oversampling	7
3.5.3	ADASYN Oversampling	7
3.6	Special Czech, Slovak, and German letters	7
3.7	Acronyms	7
3.8	Figures	7
3.9	Tables	9
3.10	Boxes	10
3.11	Theorems, Definitions,	10
3.12	Equations	11
3.12.1	Nonumbered Equations	11
3.12.2	Numbered Equations	11
3.12.3	Matrix Equations	11
3.13	Cross-references	11
3.14	Source codes	12
3.15	Paragraphs	12
4	Application of Machine Learning Algorithms	13
4.1	Repository and Environment Structure	14
4.2	Dataset Description	14
4.3	Data Exploration	14
4.3.1	Default Distribution	14
4.3.2	Categorical Features	14
4.3.3	Continuous Features	14
4.3.4	Missing Values Analysis	14

4.3.5	Association Analysis	14
4.4	Data Preprocessing	14
4.4.1	Data Split and ADASYN Oversampling	14
4.4.2	Optimal Binning and Weight-of-Evidence Encoding	14
4.5	Modelling	14
4.5.1	Feature Selection	14
4.5.2	Model Selection	14
4.5.3	Model Building	14
4.6	Model Evaluation	14
4.6.1	Confusion Matrix	14
4.6.2	Metrics Scores	14
4.6.3	ROC Curve	14
4.6.4	Learning Curve	14
4.6.5	SHAP Values	14
4.7	Machine Learning Deployment	14
4.7.1	Final Model Building	14
4.7.2	Web Application	14
4.8	Itemization and Environments	14
5	Deployment of Web Application	17
5.1	Frequently made mistakes	17
6	Title of Chapter Six	19
6.1	Useful Hints	19
7	Conclusion	22
	Bibliography	I
	A Title of Appendix A	I
	B Project's website	II

List of Tables

3.1	Calibration table	9
-----	-----------------------------	---

List of Figures

3.1	Market equilibrium	9
3.2	Boxy’s example	10

Acronyms

ML	Machine Learning
PD	Probability of Default
AUC	Area Under the Curve
LR	Logistic Regression
RF	Random Forest
GB	Gradient Boosting
MLP	Multi-Layer Perceptron
DT	Decision Tree

Chapter 1

Introduction

This document serves two purposes. First, it is a template and example for a master's thesis. Second, the text in all sections contains some useful information on structuring and writing your thesis.

The introduction should consist of three parts (as paragraphs, not to be structured into multiple headings): The first part deals with the background of the work and describes the field of research. It should also elaborate on the general problem statement and the relevance. The second part should describe the focus of the thesis, typically the paragraph starts with a phrase like “The objective of this thesis is ...” The last part should describe the structure of the thesis, for instance in the following manner. The thesis is structured as follows: Chapter 2 cites some formal requirements of the faculty and the frequently asked questions about the template, Chapter 3 gives some hints on basic formatting features and covers also acronyms, figures, boxes and tables. Chapter 4 gives a recommendation on the usage of hyphens in English language in \LaTeX and explains how to use the `itemize` and `quote` environments and shows a few `enumerate`-based environments. Chapter 5 presents a checklist of common mistakes to avoid. Chapter 6 contains numerous hints. Chapter 7 summarizes our findings.

Chapter 2

Credit Risk Modelling

2.1 Formal requirements of master's thesis

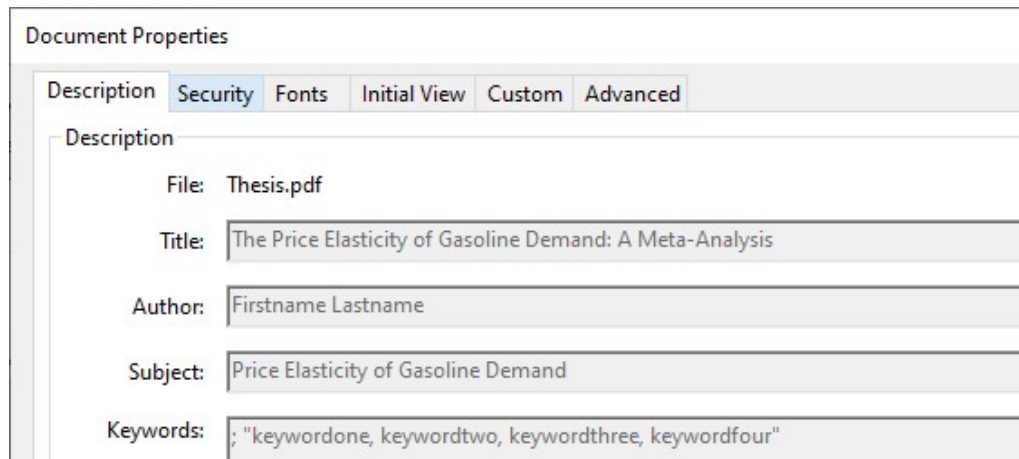
According to Dean's Provision no. 18/2017:

- The minimum extent of master's thesis is 60 standard pages (108 thousand characters including spaces) of the text itself, i.e. without an abstract and appendices and a list of literature. In case the master's thesis is written in English, its minimum extent is 50 standard pages (90 thousand characters including spaces) without an abstract and appendices and a list of literature. When writing a standard text document, the minimum requirement is 60 characters per line and 30 lines per page, i.e. 1,800 characters per page (the so-called standard page). Font size, page layout, margins, and line spacing need to be customized.
- Generally, a standard form of the page of the final thesis applies the fonts of 12 points, the gaps between the paragraphs are recommended to be of the size of 6 points. Notes and footnotes can be written in a 10-point font. The text is aligned on both sides (aligned to a block). Electronic version of the thesis will be entered by a student/applicant for a state examination through the SIS website interface in the archive format of PDF/A version 1.3 or higher. Further details are stipulated by the rector's provision.
- The master's thesis is submitted in the accreditation language of the respective follow-up Master's study program.

Note that due to GDPR, the thesis cannot include any personal information (phone, e-mail) or signatures (neither of the author nor of the supervisor).

2.2 Template adjustments and meta-data

Read README.txt to get a jist of how the template works and how to adjust styles. You can change the properties of your pdf file, such as title, author, keywords, or publisher



The image shows a screenshot of the 'Document Properties' dialog box in a PDF viewer, specifically the 'Security' tab. The dialog has a title bar 'Document Properties' and a tab bar with 'Description', 'Security', 'Fonts', 'Initial View', 'Custom', and 'Advanced'. The 'Description' tab is selected, showing a tree view with 'Description' expanded. Below the tree, there are five fields: 'File:' with the value 'Thesis.pdf', 'Title:' with 'The Price Elasticity of Gasoline Demand: A Meta-Analysis', 'Author:' with 'Firstname Lastname', 'Subject:' with 'Price Elasticity of Gasoline Demand', and 'Keywords:' with '; "keywordone, keywordtwo, keywordthree, keywordfour"'. Each field has a text input box.

Document Properties	
Description Security Fonts Initial View Custom Advanced	
Description	
File:	Thesis.pdf
Title:	The Price Elasticity of Gasoline Demand: A Meta-Analysis
Author:	Firstname Lastname
Subject:	Price Elasticity of Gasoline Demand
Keywords:	; "keywordone, keywordtwo, keywordthree, keywordfour"

in **Thesis.xmpdata** file. The file is editable in any text editor.

Chapter 3

Machine Learning

3.1 Terminology

3.2 Algorithms

3.2.1 Logistic Regression

3.2.2 Decision Tree

3.2.3 Logistic Regression

3.2.4 Naive Bayes

3.2.5 K-Nearest Neighbors

3.2.6 Random Forest

3.2.7 Gradient Boosting

3.2.8 Support Vector Machine

3.2.9 Neural Networks

3.3 Evaluation Metrics

3.3.1 Confusion Matrix

3.3.2 Accuracy

3.3.3 Recall

3.3.4 Precision

3.3.5 F1 Score

3.3.6 AUC

$$AUC = \int_0^1 TPR(FPR^{-1}(x)) dx \quad (3.1)$$

3.3.7 Kolmogorov-Smirnov

$$KS = \max_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \quad (3.2)$$

3.3.8 Somer's D

$$SD = \frac{\tau(X, Y)}{\sqrt{\tau(X, X) \tau(Y, Y)}} \quad (3.3)$$

3.3.9 Matthews Correlation Coefficient

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (3.4)$$

3.3.10 Brier Score Loss

$$BSL = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (3.5)$$

3.3.11 Jaccard Score

$$JC = \frac{1}{n} \quad (3.6)$$

3.3.12 Zero-One Loss

$$ZOL = \frac{1}{n} \sum_i \delta_{y_i \neq \hat{y}_i}, \text{ where } \delta_{y_i \neq \hat{y}_i} = \begin{cases} 1, & \text{if } y_i \neq \hat{y}_i. \\ 0, & \text{otherwise.} \end{cases} \quad (3.7)$$

3.4 Hyperparameter Tuning

dfgdfgf

3.4.1 Grid Search

3.4.2 Random Search

3.4.3 Bayesian Optimization

3.5 Imbalanced Class Distribution

3.5.1 Random Oversampling

3.5.2 SMOTE Oversampling

3.5.3 ADASYN Oversampling

dfgdfgf

Text text text text text text text text text text text text text text.
Text text text text text text text text text text. Text text text text text text
text text text text text text text text text. Text text ?.

Text text text text text text text text text text text text text text.
Text text text text text text text (see, *inter alia*, ?, pg. 10).

3.6 Special Czech, Slovak, and German letters

ů, á, š, ď, ě, ř, ô, ß, ö

3.7 Acronyms

Text text text text text text text text text text text text.
Text text text text text text text text text text. Text text text text text.
Politicians usually like inward **FDI!** (**FDI!**) and an **MNC!** (**MNC!**) appreciates
FDI! subsidies. Are **MNC!**s greedy?

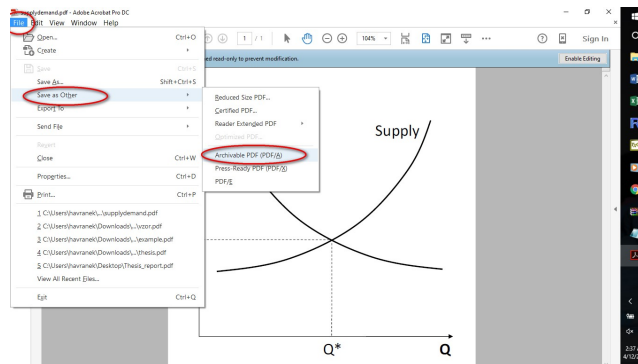
3.8 Figures

To achieve compatibility with PDF/A 2u, your file must not include links to external fonts, audio, video, or scripts. On the other hand, your file must declare each color environment you use, it must include all the pictures/figures either in jpeg or PDF/A 2u format, used fonts compliant under Unicode (your

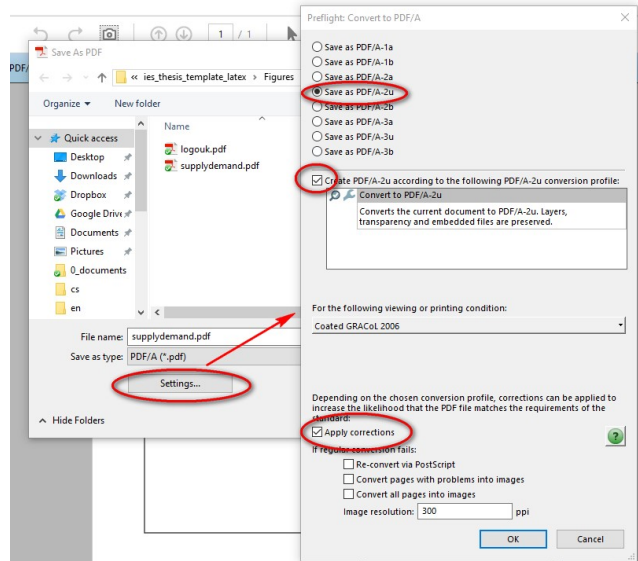
file cannot use any external fonts), and it must include meta-data in XMP format.

Most troubleshooting comes from the conversion of figures to compliant formats. You can convert from simple PDF using Adobe Acrobat:

- Select File » Save as Other » Archivable PDF (PDF/A)



- Save as PDF/A-2u:



But most of the vector graphics gets distorted to lower quality in Adobe (like pictures in pdfs generated from Stata, unless jpeg is sufficient for you). You can also use GhostScript, the conversion tool is provided by courtesy of the Faculty of Mathematics and Physics at

<https://kam.mff.cuni.cz/pdfix/>

Text text text text text text text text text. Text text text text text text text text text text.

3.10 Boxes

Text text text text text text text text text text text. Text text text text text text text text. Text text text text text text text. Text text text text text text text text text text. Let us make a box:

Figure 3.2: Boxy’s example

- Welcome to Boxy paragraph. We sincerely hope you will all enjoy the show.
- Welcome to Boxy paragraph. We sincerely hope you will all enjoy the show.
- Welcome to Boxy paragraph. We sincerely hope you will all enjoy the show.

Source: ?

Text text text text text text text text text text text. Text text text text text text text text. Text text text text text text text. Text text text text text text text text text text. Text text text text text text text text text text.

3.11 Theorems, Definitions, ...

Definition 3.1 (My original definition). This is a definition.

Assumption 3.1 (My realistic assumption). This is an assumption.

Proposition 3.1 (My clever proposition). *This is a proposition.*

Lemma 3.1 (My useful lemma). *This is a lemma.*

Example 3.1. This is an example.

Proof. This is a proof.

□

3.12 Equations

3.12.1 Nonnumbered Equations

Text text text text text text text text text text text text.
Text text text text text text text text text text. Text text text text text text.
Text text text text text text text text text text. Text text text text text text
text text text text.

$$U = \underbrace{\int_0^\infty \frac{1}{1-\sigma} (C^{1-\sigma} - 1) e^{-\rho t} dt}_{\text{meaning of life}}$$

3.12.2 Numbered Equations

Text text text text text text text text text text text text.
Text text text text text text text text text text. Text text text text text text.
Text text text text text text text text text text. Text text text text text text
text text text text.

$$U = \int_0^\infty \overbrace{\frac{1}{1-\sigma} (C^{1-\sigma} - 1)}^{\text{instantaneous utility}} e^{-\rho t} dt \quad (3.8)$$

3.12.3 Matrix Equations

Text text text text text text text text text text text text.
Text text text text text text text text text text. Text text text text text text.
Text text text text text text text text text text. Text text text text text text
text text text text.

$$\mathbf{A} = \mathbf{B} + \mathbf{C} \quad (3.9)$$

3.13 Cross-references

- to literature (?, pg. 10) or ?, pg. 10,
- to Figure 3.1,
- see Table 3.1,
- to Section 3.12,

- ### 3.14 Source codes

```
omega = 1;
syms zeta;
jmn = [1 2*zeta*omega omega^2];
figure(1);
    for zeta = 1E-5 : 0.2 : 1+1E-12
        G = tf(omega^2,subs([1 2*zeta*omega omega^2]));
        bode(G); hold on;
    end
legend('\zeta = 0', '\zeta = 0,2', '\zeta = 0,4', '\zeta = 0,6','');

```

3.15 Paragraphs

Text text text text text text text text text text text text. Text
text text text text text text text text. Text text text text text. Text
text text text text text text text text. Text text text text text text
text text text. Text text text text text text (?). Let us make two paragraphs:

Velit Text text text text text text text text text text text text. Text text text text text text text text text. Text text text text text text. Text text text text text text text text text text text text. Text text text text text text text text text text.

Chapter 4

Application of Machine Learning Algorithms

4.1 Repository and Environment Structure

4.2 Dataset Description

The analyzed dataset pertains to the HMEQ dataset which contains loan application information and default status of 5,960 US home equity loans. Such dataset was acquired from **Credit Risk Analytics**.

Table 4.1: Dataset columns

Columns	Description	Data type
BAD	Default status	Boolean
LOAN	Requested loan amount	numeric
MORTDUE	Loan amount due on existing mortgage	numeric
VALUE	Value of current underlying collateral property	numeric
REASON	Reason of loan application	character
JOB	Job occupancy category	character
YOJ	Years of employment at present job	numeric
DEROG	Number of derogatory public reports	numeric
DELINQ	Number of delinquent credit lines	numeric
CLAGE	Age of the oldest credit line in months	numeric
NINQ	Number of recent credit inquiries	numeric
CLNO	Number of credit lines	numeric
DEBTINC	Debt-to-income ratio	numeric

Source: Source: <http://www.creditriskanalytics.net/datasets-private2.html>

4.3 Data Exploration

4.3.1 Default Distribution

4.3.2 Categorical Features

4.3.3 Continuous Features

4.3.4 Missing Values Analysis

4.3.5 Association Analysis

4.4 Data Preprocessing

4.4.1 Data Split and ADASYN Oversampling

4.4.2 Optimal Binning and Weight-of-Evidence Encoding

4.5 Modelling

4.5.1 Feature Selection

4.5.2 Model Selection

4.5.3 Model Building

4.6 Model Evaluation

4.6.1 Confusion Matrix

4.6.2 Metrics Scores

4.6.3 ROC Curve

4.6.4 Learning Curve

4.6.5 SHAP Values

4.7 Machine Learning Deployment

4.7.1 Final Model Building

4.7.2 Web Application

4.8 Itemization and Environments

sight—requires m-dash. Text text text text text text text text text text text text text text text. Text text text text text text text text text text text text. Text text text text text text ?.

Text text text text text text text text text text text text text text text. Text text text text text text text text text text. Text text text text text text. Text text text text text text text text text text text text text text text. Text text text text text text text text text text text. Text text text text text text text. Text text text text text text text ?.

Let us describe the following animals:

Item 1 Text text text text text text text text text text text text text text text. Text text text text text text text text text text. Text text text text text text. Text text text text text text text text text text text text text text text text. Text text text text text text text text text text text text text text text. Text text text text text text text text text text. Text text text text text text text text text text.

Item 2 Text text text text text text text text text text text text text text text. Text text text text text text text text text text. Text text text text text text. Text text text text text text text text text text text text text text text text. Text text text text text text text text text text. Text text text text text text text text text text. Text text text text text text text text text text.

Text text text text text text text text text text text text text text text. Text text text text text text text text text text. Text text text text text text. Text text text text text text text text text text text text text text text. Text text text text text text text text text text text. Text text text text text text. See what Edmund Burke said about the duties of a Member of Parliament (Speech To The Electors Of Bristol At The Conclusion Of The Poll, November 3, 1774):

It ought to be the happiness and glory of a representative to live in the strictest union, the closest correspondence, and the most unre-served communication with his constituents. Their wishes ought to have great weight with him; their opinion, high respect; their busi-ness, unremitted attention. It is his duty to sacrifice his repose, his pleasures, his satisfactions, to theirs; and above all, ever, and in all cases, to prefer their interest to his own. But his unbiased opin-ion, his mature judgment, his enlightened conscience, he ought not to sacrifice to you, to any man, or to any set of men living. These he does not derive from your pleasure; no, nor from the law and

Chapter 5

Deployment of Web Application

5.1 Frequently made mistakes

The following checklist should help in avoiding some frequently made mistakes, if any of the following propositions apply for your thesis, there is a problem:

- You have citations in your abstract.
- The introduction does not cover the three parts as described in Chapter 1.
- The introduction contains subheadings.
- You described different aspects than promised in the title.
- You copied some parts of the text from other work without proper referencing and citing.
- You used automatic translation tools to produce text by translating it from another language.
- Your thesis contains many typos and grammatical errors. (Use an electronic spell checker. Please!)
- You used color in your figures and refer to the “blue” line (assume that your readers use a monochrome printer).
- You mainly used websites and other unrefereed material as your sources or you used Wikipedia as your source.

- You refer to something in your conclusion which you have not mentioned before.
- Some forenames in the references are abbreviated, some not.
- Some references miss a publishing date.

Chapter 6

Title of Chapter Six

6.1 Useful Hints

If you write in English, you might find the following hint useful: The indefinite article *a* is used as an before a vowel sound—for example an apple, an hour, an unusual thing, an MNC! (because the acronym is pronounced Em-En-See). Before a consonant sound represented by a vowel letter *a* is usual—for example a one, a unique thing, a historic chance. Few more tips to follow:

- Don't give orders—don't write in the imperative mood—unless you are training to be a teacher.
- Avoid the use of questions. You may know the answer: does your reader? It's much safer to tell her, or him.
- Do not become entangled in the problems of 'sexist' language. It is much easier to write in the plural. "Students should check their work" is good English. "A student should check—" is also good English, but now the problems begin: "—her work?" "—his work?" Which? You can write "his or her," but that seems clumsy. Stick to the plural.
- If you must refer to yourself, use the third person such as "The present writer would recommend that . . ." may be useful.
- Use the full forms of words and phrases, not contractions like "he's," "don't," etc. Keep the apostrophe to indicate possession—

and use it correctly. Academics really sneer at students who use the “Greengrocer’s apostrophe.”

- Do not despise short, workmanlike, and effective plain English words. If they mean what you want to say. Accurately.
- Avoid the use of humor in academic writing—unless you are very sure of yourself.
- Even when you are not being funny, avoid the use of irony or sarcasm.
- Paragraphs in academic English should contain more than one sentence. (Short paragraphs look as if you are writing for a tabloid newspaper—or a simple Template!) I guess that the average academic book runs to two or three paragraphs per page. Look at the books in your subject, and get a feel for how long your own paragraphs should be when you are imitating the academic style.
- Develop an academic vocabulary. The ‘long words’ you learn in the course of your studies are long usually because they have more precise meanings than their less formal equivalents. They are therefore better when you want to be accurate. (Also they allow you to sound like someone who deserves a degree.)
- Use as few words as you can; but use enough words to express your meaning as fully as you can. Your judgment of what is appropriate here is part of what you should learn throughout your course.
- Avoid lazy words such as “nice”. It is usually better to say “acquire” or “obtain” than “get;” and it may be better, if you mean “through the use of money,” to say “purchase” or—better still—“buy.”
- A short word like “buy” is better than a long one like “purchase”—unless the long one is more accurate. A “statutory instrument” is better than a “rule”—to a lawyer, at any rate.

-
- Proof-read with care. Ask someone else to help—you may be too close to your work to be able to see your mistakes.
 - If in doubt, choose the more formal, or possibly just the more old-fashioned, of two words. For example, say quotation rather than quote whenever you mean the use of somebody else's words.
 - You will often sound more academic if you include doubts in your work—and qualifications. Within the scope of this thesis, the current writer cannot hope to cover all the possible implications of the question.
 - In this context, the use of litotes sounds very academic. This is the construction where a writer uses a negative with a negative adjective, e.g. it is not unlikely that ... This does not mean the same as it is probable that ... It has a shade of meaning and qualification that can be useful to academic writers.

Chapter 7

Conclusion

The conclusion should briefly summarize the problem statement and the general content of the work and the emphasize on the main contribution of the work.

When writing the conclusion keep in mind that some readers may not have gone through the whole thesis, but have jumped directly to the conclusion after having read the abstract in order the decide on the personal relevance of the thesis. Therefore, the conclusion should be self contained, which means that a reader should be able to understand the essence of the conclusion without having to read the whole thesis.

The conclusion typically ends with an outlook that describes possible extensions of the presented approaches and of planned future work.

Appendix B

Project's website

You can create a special website for your project which contains empirical data and MatLab/R/Stata source codes, see meta-analysis.cz/sigma, for example. Stating in your thesis that the data and source codes are available upon request is enough but please, have them prepared for such requests. The faculty does not allow enclosed DVD.

- File 1: Master's thesis
- File 2: Empirical data
- File 3: Source codes