The Government of the Russian Federation

Federal State Autonomous Educational Institution of Higher Professional Education
“National Research University “Higher School of Economics”

Faculty of business and management
School of business Informatics

Program of Discipline
“Predictive Modeling”

for Master training direction 38.04.05 “Business Informatics”

The authors of the program:
Maria S. Poptsova

Approved at the meeting of the Department of
information and business in the sphere of information technologies «____»_____ 2016

Head of Department, Svetlana V. Maltseva _____

Approved by the Academic Council of master’s program «Big Data Systems»
«____»_____ 2016.
Academic Supervisor, Svetlana V. Maltseva _____

Moscow, 2016

1. Course Description

a. Title of the course

Predictive Modeling

b. Pre-requisites

This program of an academic discipline establishes minimum requirements for knowledge and skills of the student and determines the content and types of studies and reports.
The program is designed for teachers, leading this discipline, teaching assistants and students directions 38.04.05 "Business Informatics" Master training, students in the master's program "Business Informatics".
The program is developed in accordance with:

- the educational standards of the Federal State Autonomous Educational Institution of Higher Professional Education "National Research University "Higher School of Economics", the level of training: Master, approved by 26.06.2011;
- working curriculum of the University towards 38.04.05 "Business Informatics" Master training for master's program «Big Data Systems», approved in 2013

c. Course type

Elective

d. Abstract

Predictive modeling is the process of developing a model so that it becomes possible to understand and quantify the model prediction accuracy in the future. A developed model helps to better predict future outcomes for the variable of interest by exploring its relationships with other variables from the recorded data. Due to technological advances organizations have nowadays collected enormous amount of raw data, which needs to be processed to extract actionable information and knowledge. Thus predictive modeling can be used in many business applications including customer relationship management, risk management, and process management.

The course introduces the theory and practice of predictive modeling in a data- rich modern science and business environment. Predictive models such as linear and nonlinear regression, classification models and decision trees, association analysis, and link analysis will be studied. Workshops, which follow the lectures, seek to empower students with the practical skills in predictive modeling software tools, packages and applications. Many case studies of predictive models from different areas will be considered.

2. Learning Objectives

- Developing theoretical knowledge and practical basic skills for understanding the process and key parts of predictive model building, understanding types of predictive models and the main caveats in their creation.

- Developing practical skills for data pre-processing, predictor selection, model tuning, measuring predictive model performance, finding factors affecting model performance.

- Developing practical applications in different areas of business and science, from conceptual modeling through to implementation in computer program.

3. Learning Outcomes

Having taken the course, the students will

- know the theory of the process and components of predictive modeling, types of predictive models, key steps of model creation, such as data-preprocessing, model construction and assessment of model performance.
- know various practical applications of predictive modeling from science to business.
- acquire the skills to use R functions from different R packages to apply different types of models such as linear and nonlinear regression models, linear and nonlinear classification models, regression trees and rule-based models.
- acquire the skills to use R functions from different R packages to pre-process the input data, i.e. calculate statistics, estimate skewness, apply appropriate transformation, perform PCA, find between-predictor correlations, generate dummy variables.
- acquire the skills to use R functions to measure predictor importance and model performance, use filtering methods, measure outcome error.
- apply the knowledge and tools of predictive analytics to real-life applications.

As a result of the development of the discipline the student acquires the following competences:

Competence	GEF/NRU code	Descriptors - the main features of the development (indicators of achievement results)	Forms and methods of teaching, contributing to the formation and development of competence
Ability to offer concepts, models, invent and test methods and tools of professional activity	СК-2	Demonstrates	Lectures, workshops, homework
The ability to apply the methods of system analysis and modeling to evaluate and design	ПК-13	Owns and uses	Lectures, workshops, homework
Ability to develop and apply mathematical models to justify the design decisions in the field of predictive analytics	ПК-14	Owns and uses	Lectures, workshops, homework
Ability to organize self and collective research work at the enterprise and manage it	ПК-16	Owns and uses	Lectures, workshops, homework

4. Course Plan

№	Topic name	Total hours	Classroom hours			Homework
			Lectures	Seminars	Workshops	
1	Predictive Modeling Process	18	2		4	12
2	Data Pre-processing	18	2		4	12
3	Over-fitting and Model Tuning	18	2		4	12
4	Linear Regression Models	18	2		4	12
5	Nonlinear Regression Models	24	4		6	14
6	Linear Classification Models	18	2		4	12
7	Nonlinear Classification Models	24	4		6	14
8	Regression Trees and Rule-Based Models	18	2		4	12
9	Measuring Predictor Importance	18	2		4	12
10	Feature Selections	18	2		4	12
11	Factors Affecting Model Performance	18	2		4	12
12	Prediction in Customer Relationship Management	18	2		4	12
	Total	**228**	**28**		**52**	**148**

5. Grading System

The student should demonstrate the knowledge of sections of the discipline and the ability to present the results of homework and tests in accordance with the required competencies. Evaluation of all forms of monitoring are set on a 10-point scale.

Grade	10-point scale
Excellent	10
	9
	8
Good	7
	6
Satisfactorily	5
	4
Bad	3
	2
	1
	0

6. Guidelines for Knowledge Assessment

On the final evaluation (O_i) on a subject matter consists of ratings for:

- practical work – O_1
- exam (including oral and practical parts) – O_2

according to the formula: $O_i = 0.6*O_1 + 0.4* O_2$

7. Methods of Instruction

Presentation, cross-subject practices, collective practice work (programming and projects).

8. Course Content

1. Predictive modeling process

Key parts of predictive models. Concepts of model building. Data "spending". Data splitting. Predictors. Candidate models. Optimal model. Performance estimation. Case study: fuel economy.

Workshop: Introduction to R. Packages. Data Types and Basic Structures. Working with Rectangular Data Sets. Objects, Classes, Functions. R Packages: *AppliedPredictiveModeling*, *caret (*short for *Classification And REgression Training).*

2. Data pre-processing

Unsupervised data processing. Techniques of addition, deletion, transformation of training data set. Reduction of data skewness or outliers. Feature engineering. Feature extraction. Surrogate variables as combinations of multiple predictors. Dummy variables. Case study: cell segmentation in high-content screening.

Workshop: data segmentation, calculation of skewness statistics, Box-Cox transformation, PCA, data rotation, predictor filtering, between-predictor correlations, generation of dummy variables. R-packages: *AppliedPredictiveModeling*, *caret, corrplot, e1071,* and *lattice.*

3. Over-fitting and model tuning

Concept of over-fitting. Model tuning and model evaluation. Tuning parameters. Resampling Techniques. k-Fold cross-validation. Generalized cross-validation. Bootstrap. Case study: credit scoring.

Workshop: creating random splits of the data, data resampling (bootstrapping, k-fold cross-validation, repeated cross-validation), basic model building in R, determination of tuning parameters, between-model comparisons. R-packages*: AppliedPredictiveModeling, caret, Design, e1071, ipred and MASS.*

4. Linear regression models

Measuring performance in regression models. Linear regression. Partial least squares and penalized models. Case study: quantitative structure-activity relationship modeling.

Workshop: Linear regression models using simple least squares, robust linear regression model, resampling estimate of performance, partial least squares, penalized regression models. R packages: *elasticnet, caret, lars, MASS, pls* and *stats.*

5. Nonlinear regression models

Neural networks. Multivariate adaptive regression splines. Support vector machines. K-nearest neighbors. Case studies.

Workshop: Neural networks, multivariate adaptive regression splines, support vector machines, k-nearest neighbors. R packages: *nnet*, *neural*, and *RSNNS.*

6. Linear classification models

Measuring performance in classification models. Sensitivity and specificity. Receiver operating characteristic curves. Logistic regression. Linear discriminant analysis. Partial least squares discriminant analysis. Penalized models. Nearest shrunken centroids.

Workshop: sensitivity and specificity, receiver operating characteristic curves, lift charts, calibrating probabilities, logistic regression, linear discriminant analysis, partial least squares discriminant analysis, penalized models, nearest shrunken centroids. R packages: *AppliedPredictiveModeling, caret, klaR, MASS, pROC*, *glmnet, pamr, pls, rms, sparseLDA,* and *subselect*.

7. Nonlinear classification models

Nonlinear discriminant analysis. Neural networks. Flexible discriminant analysis. Support vector machines. K-nearest neighbors. Na¨ive Bayes. Case studies.

Workshop: nonlinear discriminant analysis, neural networks, flexible discriminant analysis, support vector machines, k-nearest neighbors, na¨ive Bayes. R packages: *caret, earth, kernlab, klaR, MASS, mda, nnet,* and *rrcov*.

8. Regression trees and rule-based models

Basic regression trees and regression model trees. Basic classification trees. Bagged trees. Random forests. Boosting. Cubist. Case studies.

Workshop: single trees, model trees, bagged trees, random forest, boosted trees, cubist, classification trees, rules. R packages: *caret, Cubist, gbm, ipred, party, partykit, randomForest, rpart, RWeka, randomForest*, *C50, pROC*.

9. Measuring predictor importance

Quantification of the relationship between predictors and an outcome. Numeric Outcomes. Categorical Outcomes. Relief algorithm. Model-based importance scores.

Workshop: results visualization, root mean squared error *(*RMSE*)*, square of the correlation coefficient R^2, Spearman's rank correlation, numeric outcomes, categorical outcomes, model-based importance scores. R packages: *AppliedPredictiveModeling, caret, COREarn, minerva, pROC, randomForest, stats.*

10. Feature selections

Non-informative predictors. Reducing the number of predictors. Wrapper methods. Filter methods. Selection bias. Genetic algorithms. Case study.

Workshop: forward, backward, and stepwise selection, recursive feature elimination, filter methods. R packages: *AppliedPredictiveModeling, caret, klaR, leaps, MASS, pROC, rms,* and *stats.*

11 Factors affecting model performance

Type III errors. Measurement error in the outcome. Measurement error in the predictors. Discretizing continuous outcomes. When should one trust your model's prediction? Impact of a large sample.

Workshop: similarity algorithm, model tuning and computation of variable importance, predictors' permutation, model tuning on the newly created classification data. R packages: *caret*.

12. Prediction in customer relationship management

Practical applications of predictive modeling. Who will be responding to new advertisement? What customers are most likely to be default paying loans? What transactions are most likely to be fraudulent? What combinations of products are customers most likely to purchase at the same time?

Workshop: Case study of customer relationship prediction modeling.

8. Reading List

a. Required

- Max Kuhn, Kjell Johnson. Applied Predictive Modeling. Springer, 2013.
- Eric Siegel, Thomas H. Davenport. Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die. Wiley, 1 edition, 2013
- Thomas W. Miller. Modeling Techniques in Predictive Analytics: Business Problems and Solutions with R. FT Press, 1 edition, 2013.

b. Optional

- Foster Provost, Tom Fawcett. Data Science for Business: What you need to know about data mining and data-analytic thinking. O'Reilly Media, 1 edition, 2013.
- Conrad Carlberg. Predictive Analytics: Microsoft Excel. Que Publishing, 1 edition, 2012.

9. Control Questions

1. Key parts of predictive models. Concepts of model building.
2. Data “spending”. Data splitting. Predictors.
3. Candidate models. Optimal model. Performance estimation.
4. Unsupervised data processing. Techniques of the addition, deletion, transformation of training data set. Reduction of data skewness or outliers.
5. Feature engineering. Feature extraction.
6. Surrogate variables as combinations of multiple predictors. Dummy variables.
7. Concept of over-fitting.
8. Model tuning and model evaluation. Tuning parameters.
9. Resampling Techniques. k-Fold cross-validation. Generalized cross-validation.
10. Bootstrap.
11. Measuring performance in regression models.
12. Linear regression.
13. Partial least squares and penalized models.

14. Neural networks.
15. Multivariate adaptive regression splines.
16. Support vector machines.
17. K-nearest neighbors.
18. Measuring performance in classification models.
19. Sensitivity and specificity. Receiver operating characteristic curves.
20. Logistic regression. Linear discriminant analysis.
21. Partial least squares discriminant analysis.
22. Penalized models.
23. Nearest shrunken centroids.
24. Nonlinear discriminant analysis.
25. Neural networks.
26. Flexible discriminant analysis.
27. Support vector machines.
28. K-nearest neighbors.
29. Na¨ive Bayes.
30. Basic regression trees and regression model trees.
31. Basic classification trees. Bagged trees.
32. Random forests.
33. Boosting. Cubist.
34. Measuring predictor importance. Numeric Outcomes. Categorical Outcomes.
35. Relief algorithm. Model-based importance scores.
36. Feature Selections. Non-informative predictors.
37. Wrapper methods. Filter methods. Selection bias.
38. Genetic algorithms.
39. Factors that can affect model performance. Type III Errors.
40. Measurement errors in the outcome and in the predictors. Impact of a Large Sample.

10. Technical Provisions

Computer, overhead projector (for lectures or workshops), computer class.