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

This program cannot be used by other divisions of the university and other institutions of higher education without the permission of the department - the developer of the program.
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:

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

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

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.

<table>
<thead>
<tr>
<th>Grade</th>
<th>10-point scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excellent</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>8</td>
</tr>
<tr>
<td>Good</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>6</td>
</tr>
<tr>
<td>Satisfactorily</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>4</td>
</tr>
<tr>
<td>Bad</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>

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


2. Data pre-processing


3. Over-fitting and model tuning


4. Linear regression models


5. Nonlinear regression models


6. Linear classification models


**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


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

8. Regression trees and rule-based models


**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


**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`, `CORElearn`, `minerva`, `pROC`, `randomForest`, `stats`.

10. Feature selections


11. Factors affecting model performance


**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


#### b. Optional


### 9. Control Questions

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.
7. Concept of over-fitting.
8. Model tuning and model evaluation. Tuning parameters.
12. Linear regression.
13. Partial least squares and penalized models.
15. Multivariate adaptive regression splines.
20. Logistic regression. Linear discriminant analysis.
22. Penalized models.
23. Nearest shrunken centroids.
26. Flexible discriminant analysis.
27. Support vector machines.
29. Naïve Bayes.
30. Basic regression trees and regression model trees.
32. Random forests.
35. Relief algorithm. Model-based importance scores.
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.


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