

RapidMiner Automated Model Ops, Model Deployment and Management.

Sven Van Poucke, MD, PhD

Introduction

Although enormous datasets are increasingly generated, not only in the formal health care setting, but also emanating from data streams from medical and consumer devices, wearables, patient-reported outcomes, as well as environmental, community and public health sources, explainability and interpretability continue to be a significant barrier for implementing AI/ML tools in particular when physicians need to be convinced the investment in new technology leads to a higher standard of care.

However, the tools for automated model development, model deployment and management have recently become more accessible using intuitive technology enabling collaboration on models, monitor for governance, drift and bias issues, inclusive set up alerts and integration strategy options.

This document illustrates how predictive modeling can be developed using a code-free intuitive visual approach increasing explainability and interpretability without black boxes. Next the developed models are deployed and the tools available for monitoring are visualized.

Methods

The data used in this document is based on a recent PLOSONE publication by Pengcheng Yang et al. The first dataset (<https://doi.org/10.1371/journal.pone.0226962.s002>) was used for model development (inclusive validation) using RapidMiner AutoModel tool (more information can be found here: <https://docs.rapidminer.com/latest/studio/guided/auto-model/>). The second dataset (<https://doi.org/10.1371/journal.pone.0226962.s003>) was injected for model deployment monitoring taking into account that full deployment monitoring requires more additional datasets (as explained here: <https://docs.rapidminer.com/latest/studio/guided/deployments/>)

The screenshot shows a research article from PLOS ONE. At the top left, there are icons for 'OPEN ACCESS' and 'PEER-REVIEWED'. Below these, it says 'RESEARCH ARTICLE'. The main title of the article is 'A new method for identifying the acute respiratory distress syndrome disease based on noninvasive physiological parameters'. Below the title, the authors are listed: Pengcheng Yang, Taihu Wu, Ming Yu, Feng Chen, Chunchen Wang, Jing Yuan, Jiameng Xu, Guang Zhang. A small envelope icon indicates the availability of a PDF version. At the bottom, it says 'Published: February 5, 2020 • https://doi.org/10.1371/journal.pone.0226962'.

Results

Figure 1: features (n=44) used for predictive modeling

| Result History | | ExampleSet (Select Attributes (2)) | | ExampleSet (Select Attributes (2)) | |
|----------------|------------|------------------------------------|-----------------|------------------------------------|---------------|
| Data | Open in | | Open in | | Auto Model |
| | Turbo Prep | Auto Model | Turbo Prep | Auto Model | |
| Row No. | id | Row No. | id | Row No. | id |
| 1 | att1 | 16 | rr | 17 | tv |
| 2 | subject_id | 18 | pip | 19 | plap |
| 3 | hadm_id | 20 | mv | 21 | map |
| 4 | icustay_id | 22 | peep | 23 | gcsmotor |
| 5 | pao2 | 24 | gcsverbal | 25 | gcseyes |
| 6 | spo2 | 26 | first_careunit | 27 | last_careunit |
| 7 | fio2 | 28 | age | 29 | ethnicity |
| 8 | hr | 30 | admission_ty... | 31 | gender |
| 9 | temp | 32 | height_first | 33 | height_min |
| 10 | nbps | 34 | height_max | 35 | weight_first |
| 11 | nbpd | 36 | weight_min | 37 | weight_max |
| 12 | nbpm | 38 | gcs | 39 | sf |
| 13 | abps | 40 | osi | 41 | bmi |
| 14 | abpd | 42 | tv_kg | 43 | pf |
| 15 | abpm | 44 | pfclass_two_... | | |

Figure 2: Results of modeling with standard Naive Bayes parameters.

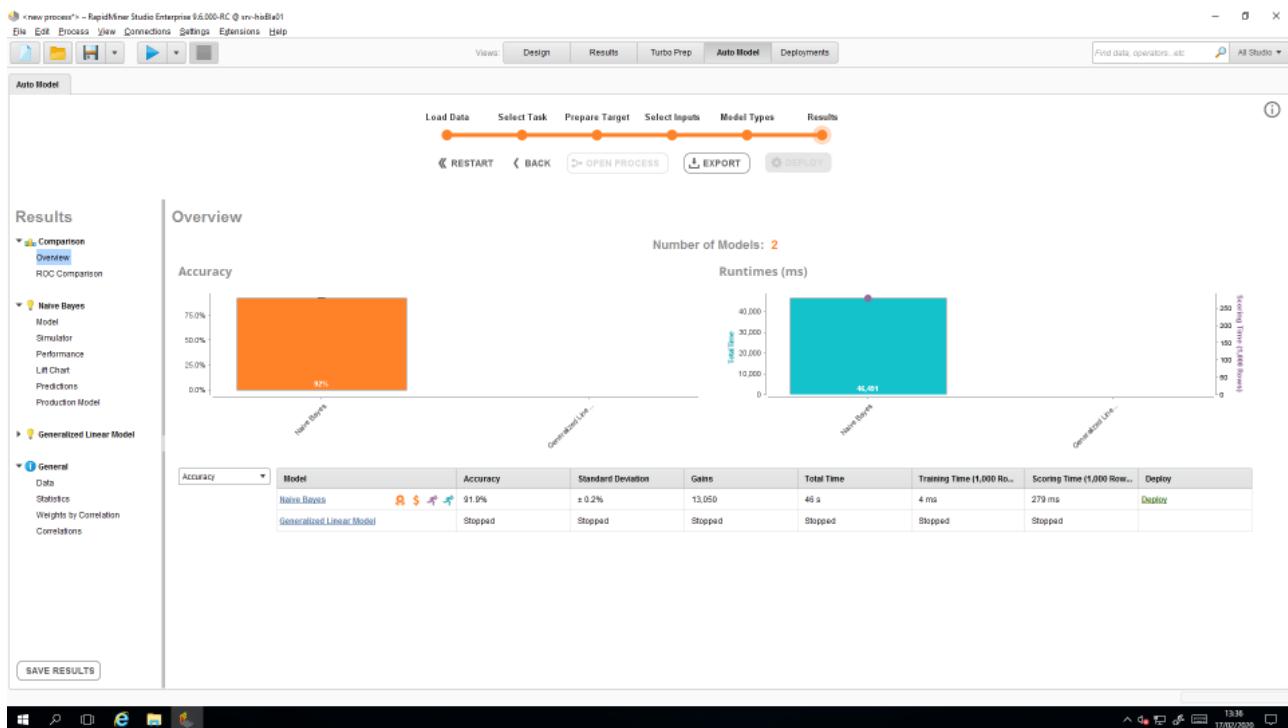


Figure 3: ROC curve of Naive Bayes.

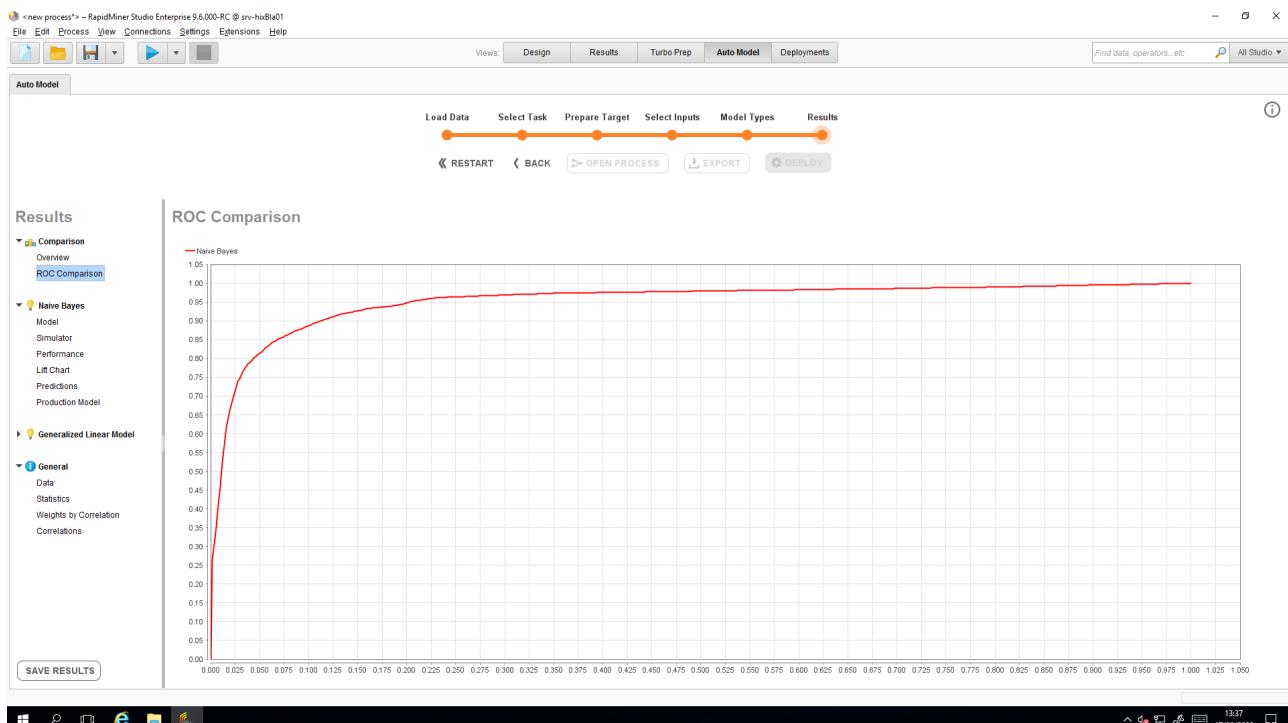


Figure 4: Weight by correlation of Naive Bayes.

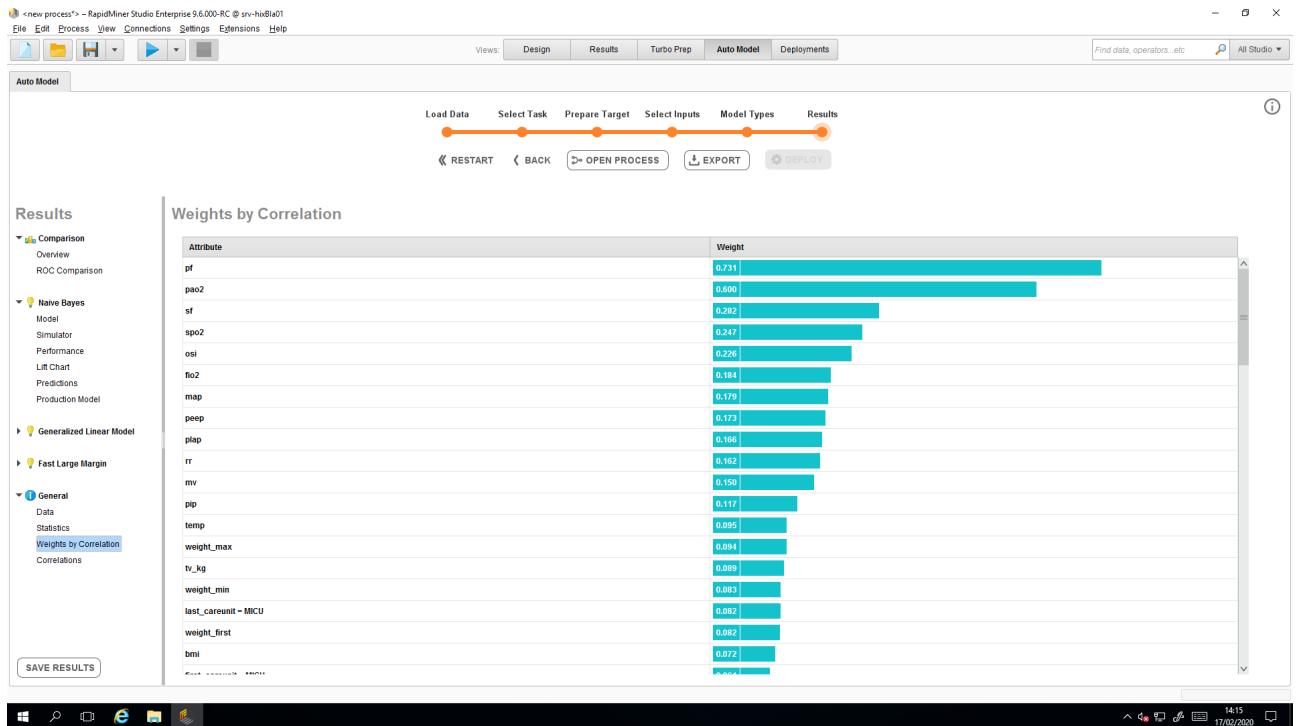


Figure 5: Performance of Naive Bayes.

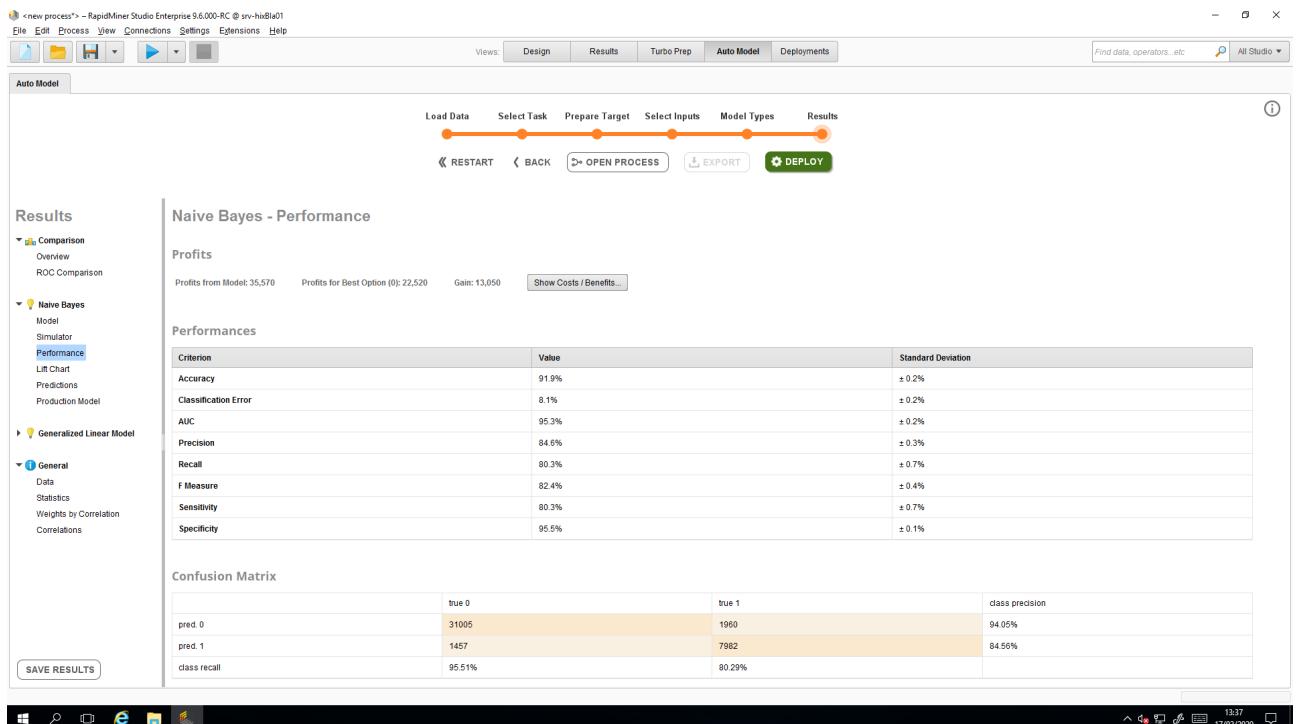


Figure 6: Data modeling by Naive Bayes, GLM, Fast Large Margin.

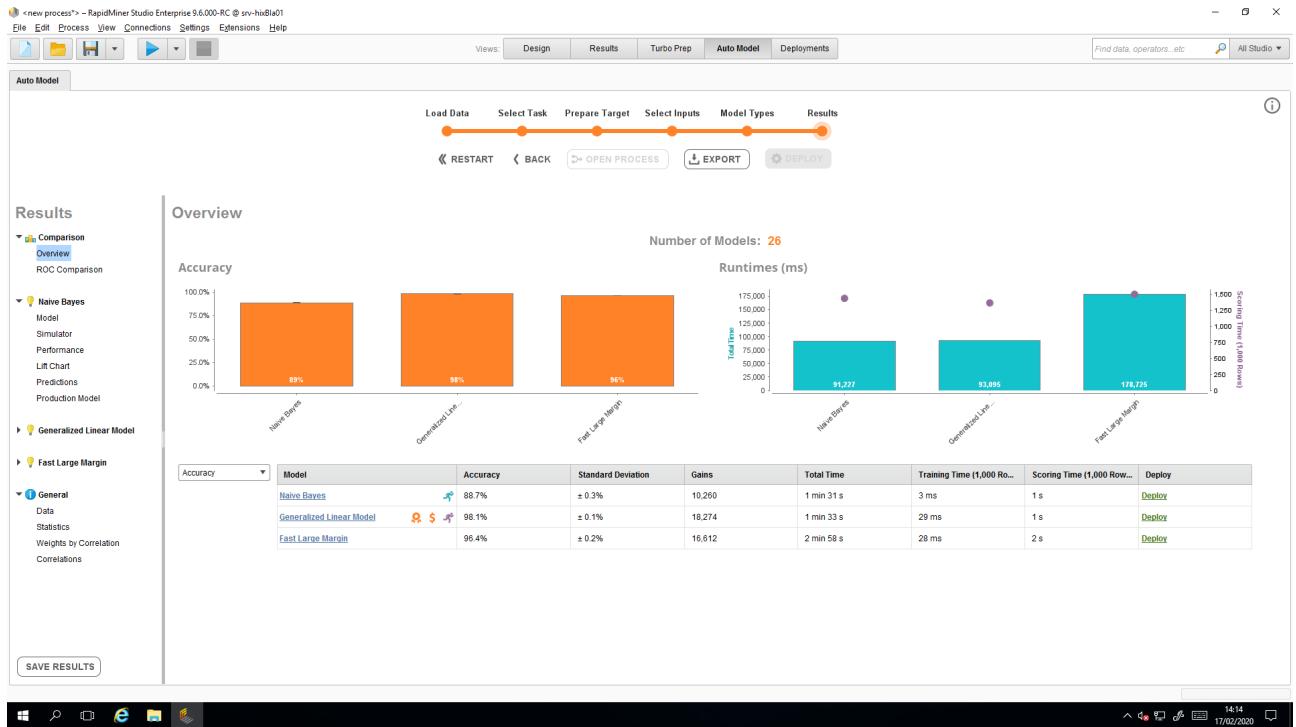


Figure 7: ROC comparison of NB, GLM, Fast Large Margin.

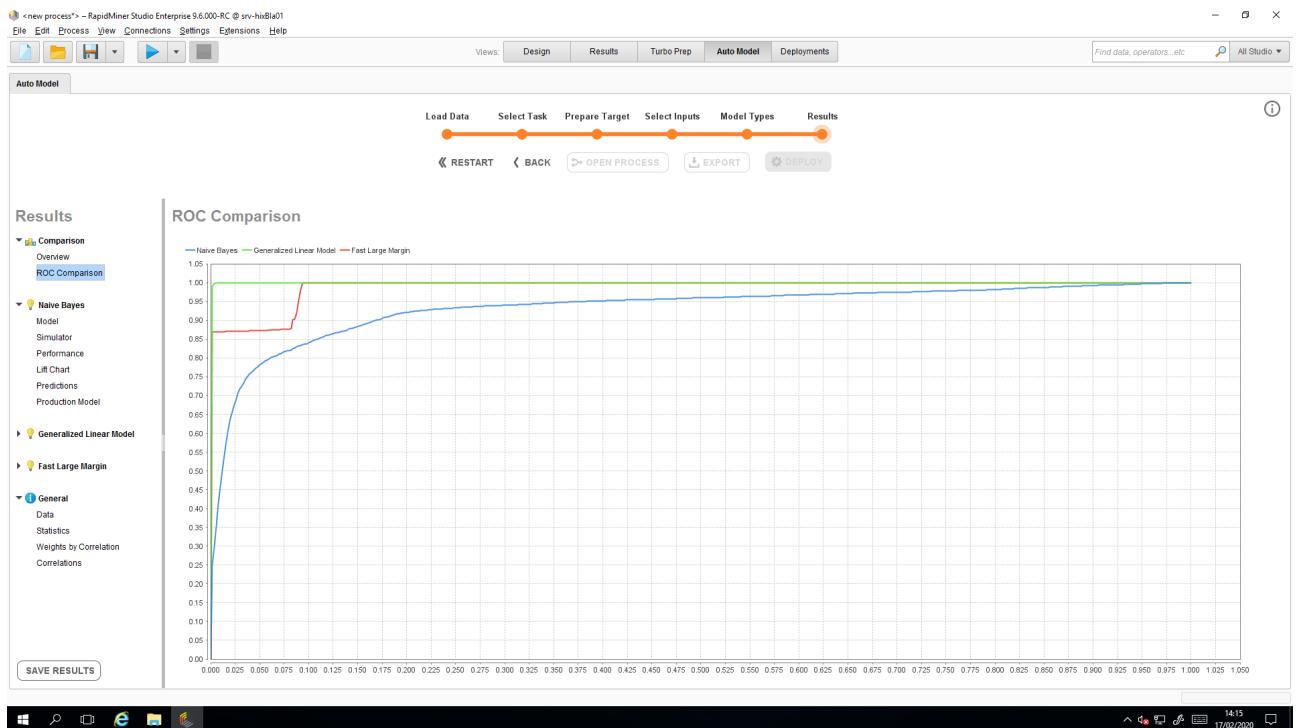


Figure 8: Correlations.

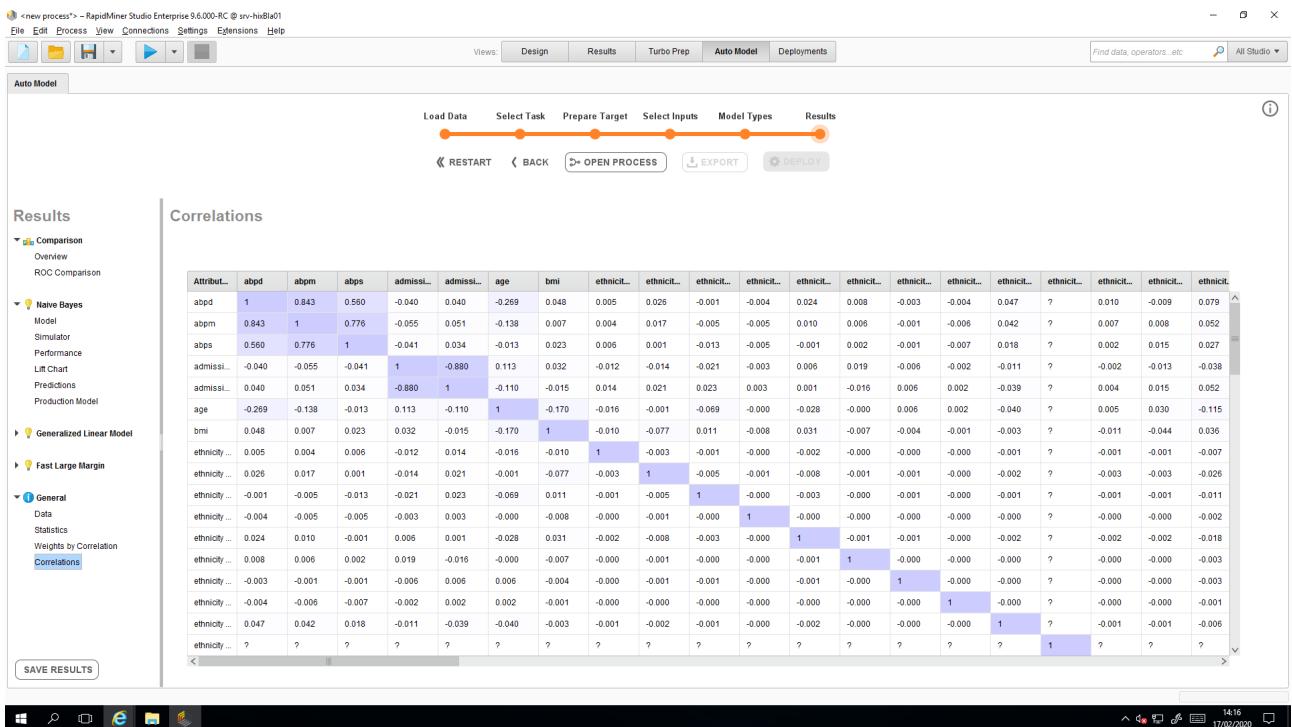


Figure 9: Deployment of NB model as use case.

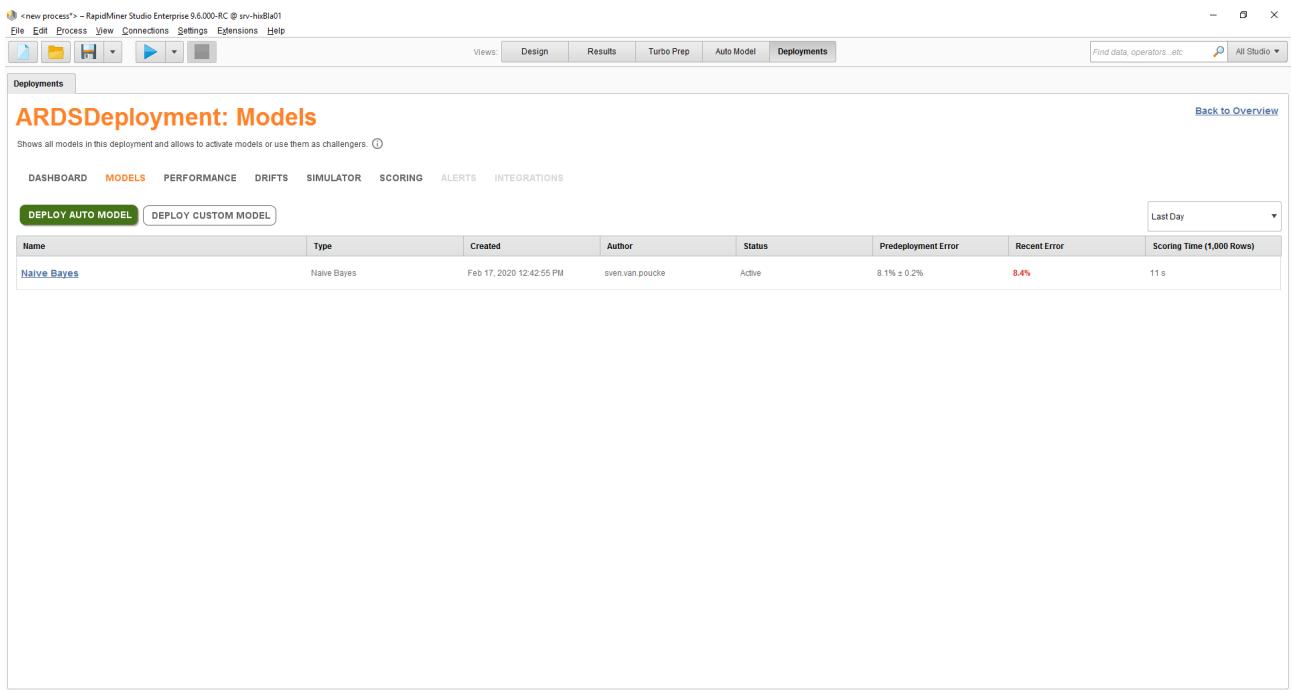


Figure 10: Deployment Dashboard.

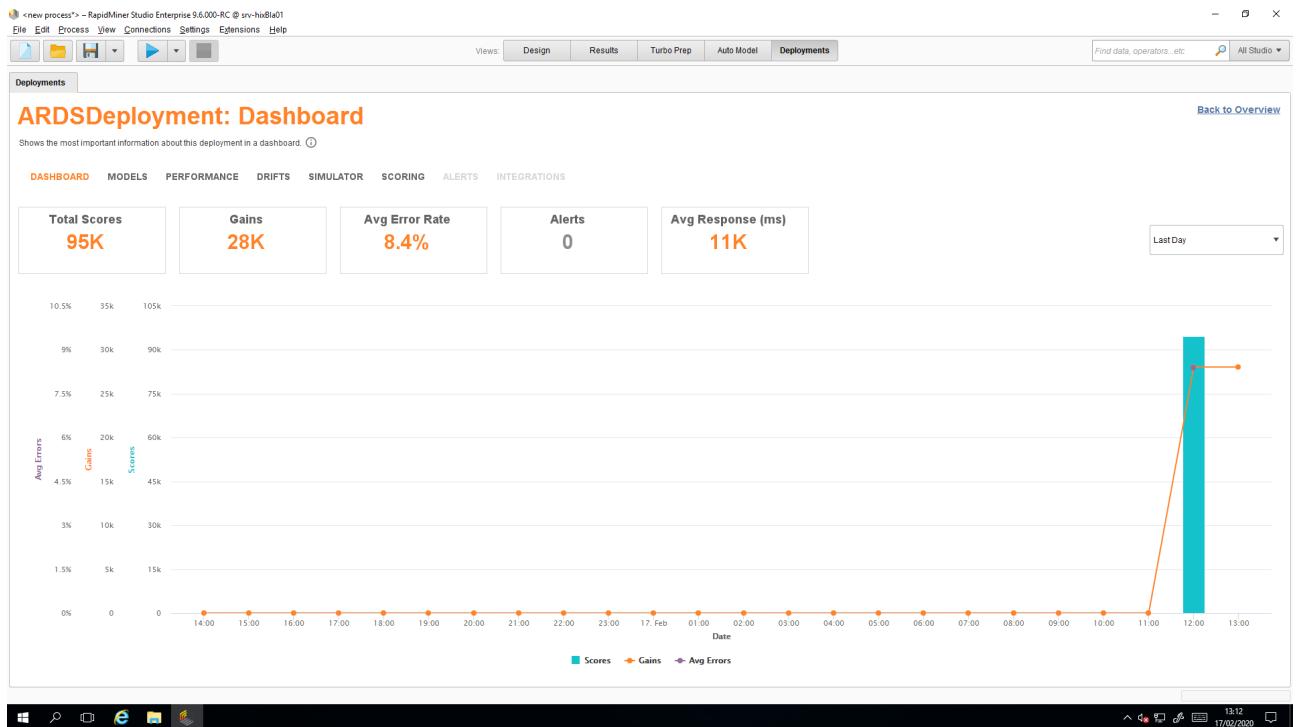


Figure 11: Performance of deployed model(s).

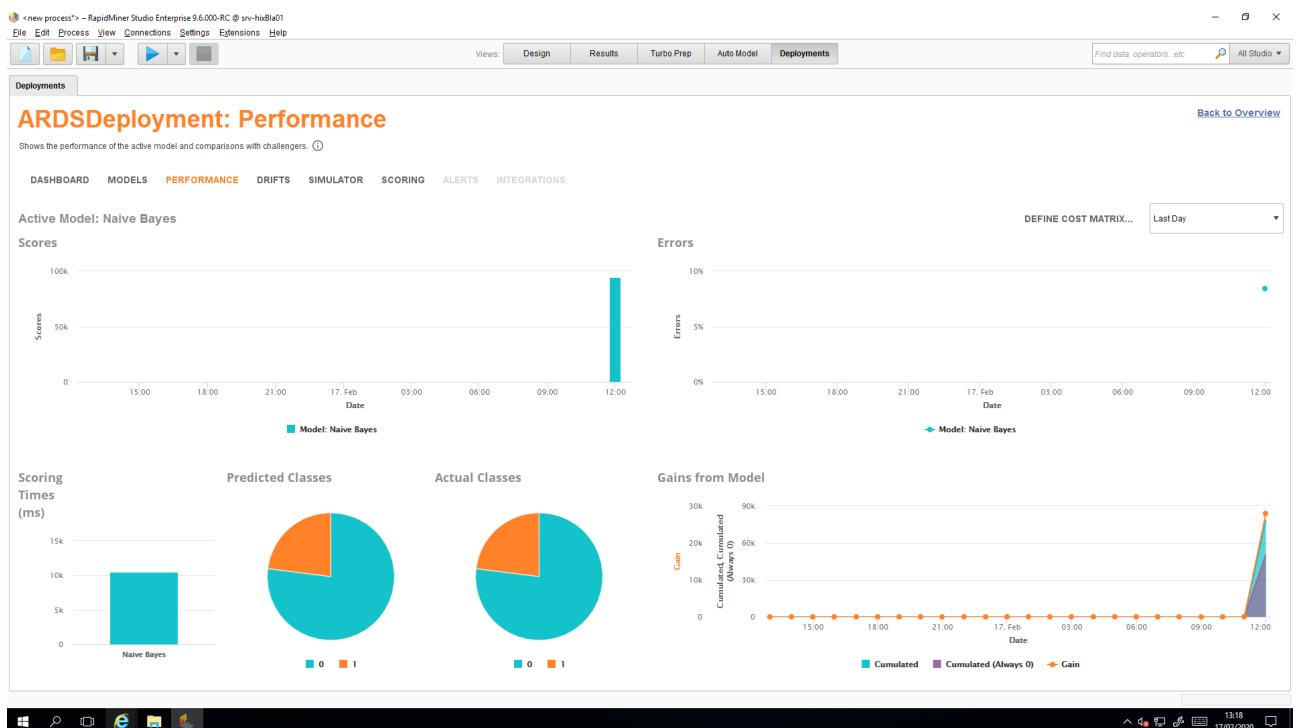


Figure 12: Drifts: highlights the columns where the data lately deviates from the training data.

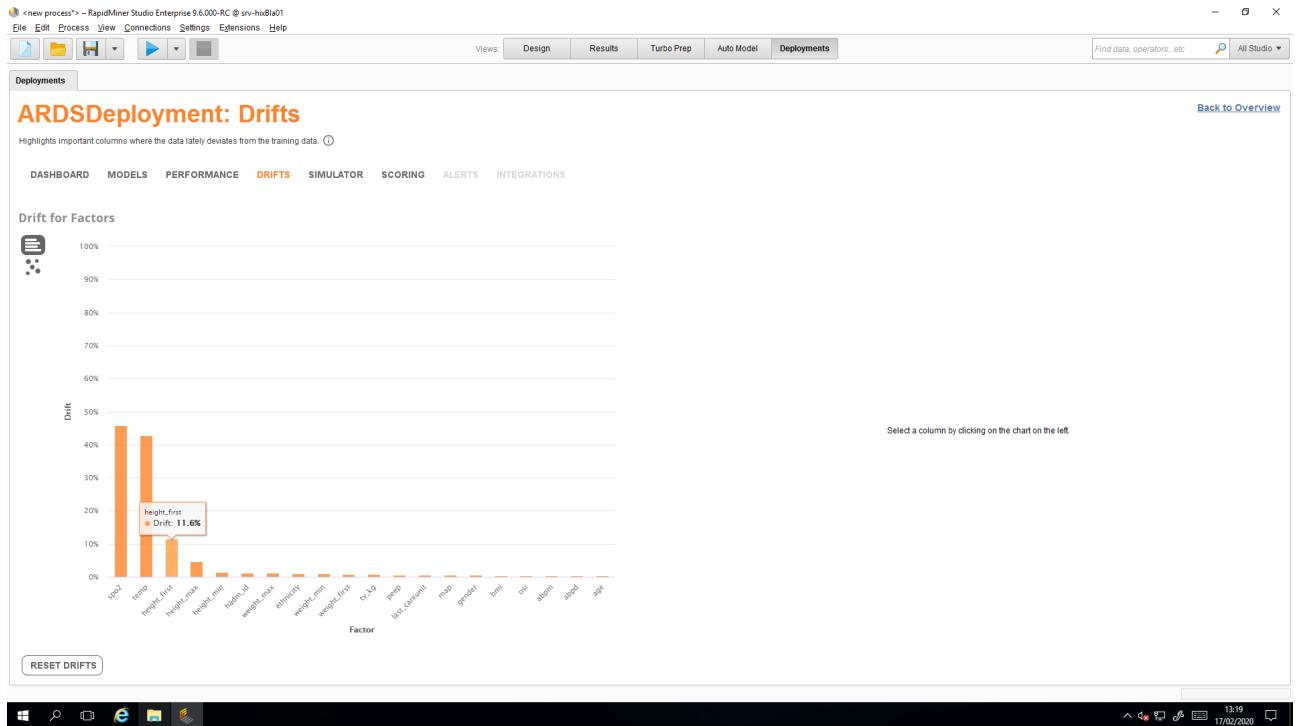
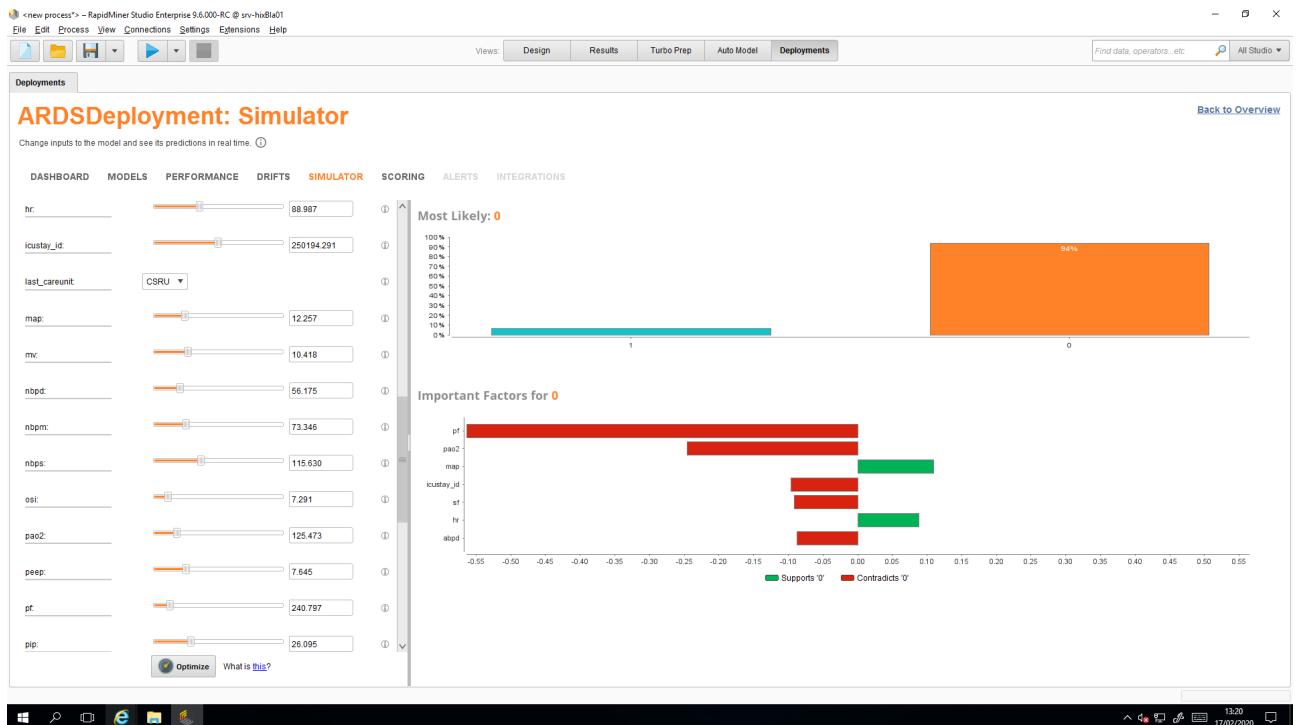


Figure 13: Deployment Simulator.



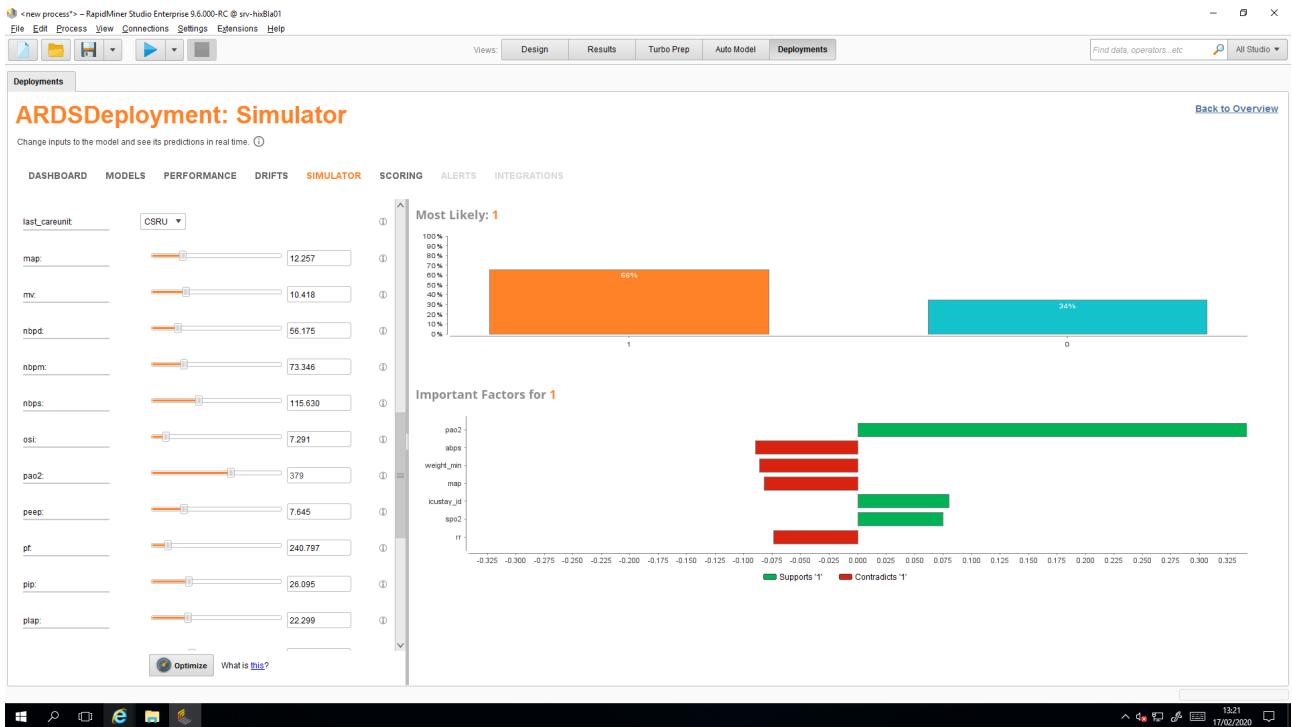


Figure 14: Input Data of Deployment Model.

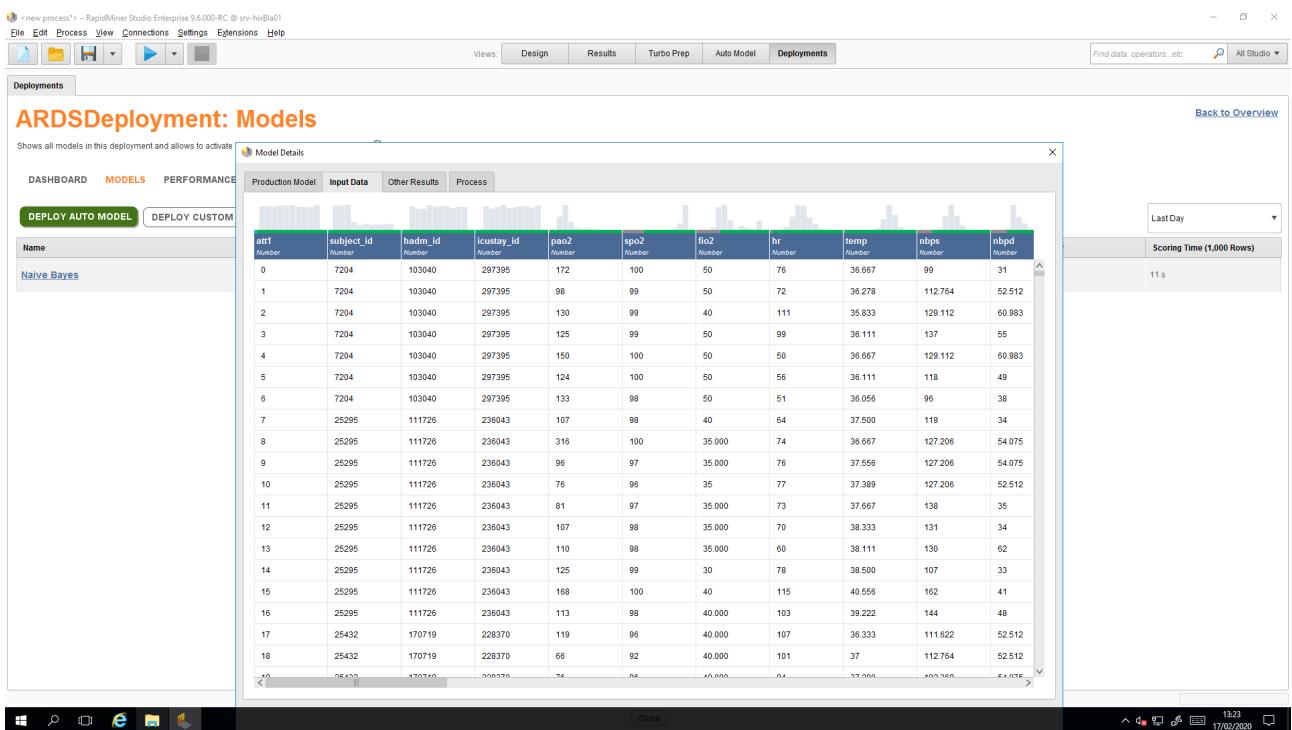


Figure 15: Production Model: Feature Details:

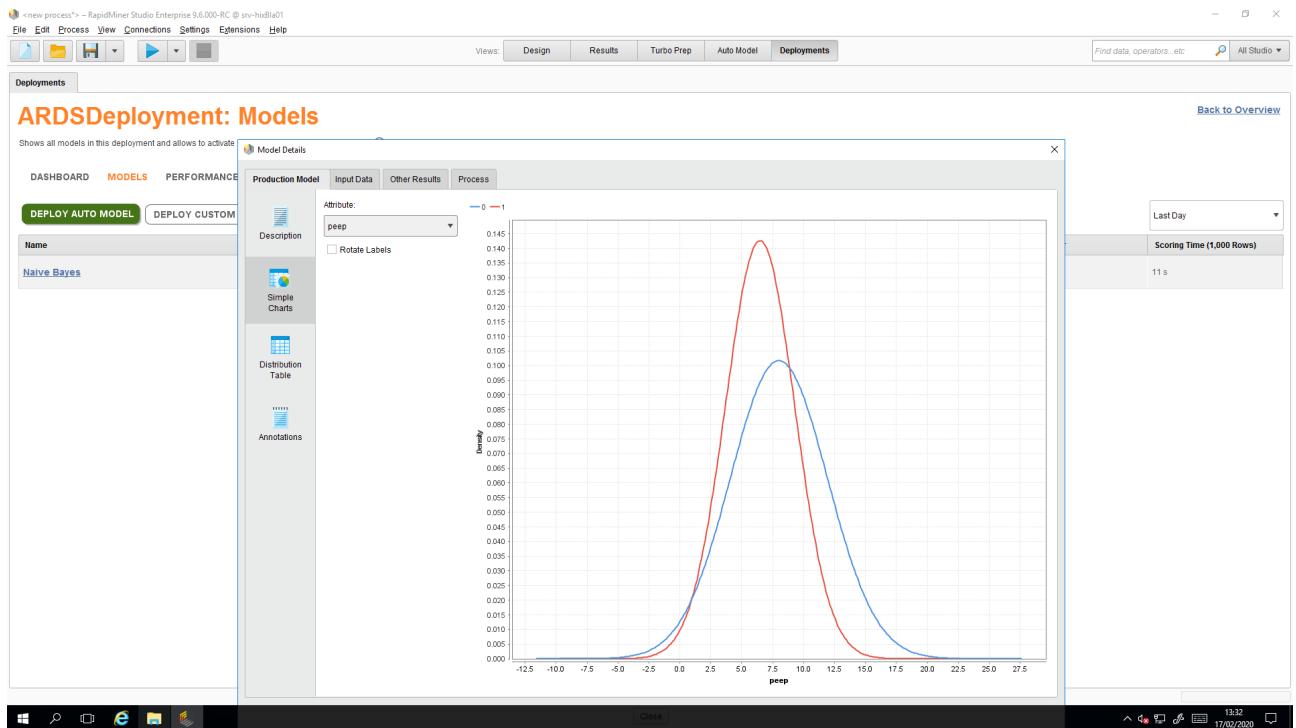


Figure 16: Lift Chart.

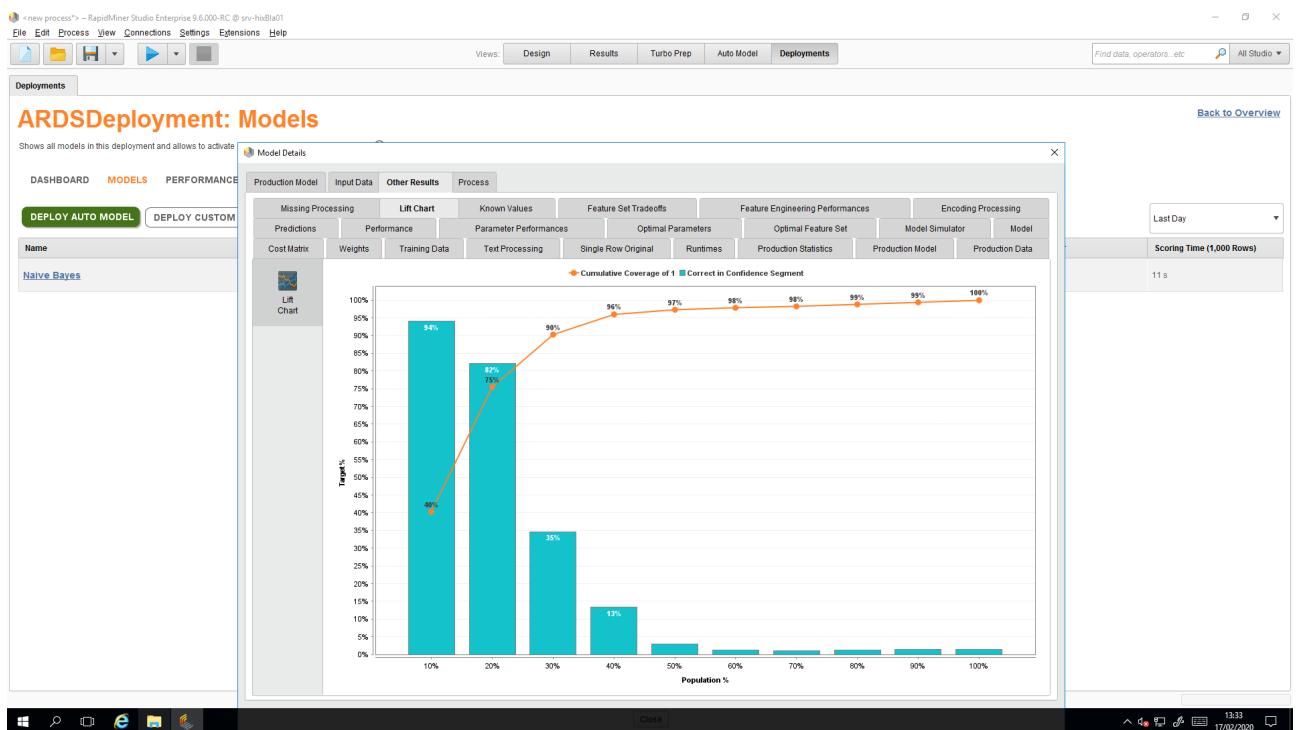


Figure 17: Predictions with confidence (0,1) and cost.

The screenshot shows the RapidMiner Studio interface with the 'Models' tab selected. A specific model named 'Naive Bayes' is highlighted. The main panel displays a table titled 'Model Details' under the 'Predictions' tab. The table includes columns for Row No., pclass_two_, prediction(p...), confidence(0), confidence(1), cost, first_careunit, last_careunit, ethnicity, admission_l..., and gender. The data shows 19 rows of predictions with various values for each column. The interface also includes tabs for 'Input Data', 'Other Results', and 'Process'. On the right side, there are filters for 'Last Day' and 'Scoring Time (1,000 Rows)', and a status bar indicating '11 s'.

Figure 18: Performance of Production Model.

The screenshot shows the RapidMiner Studio interface with the 'Models' tab selected. The same 'Naive Bayes' model is selected. The main panel now displays the 'Performance' tab of the 'Model Details' dialog. It shows a table view for accuracy, which is highlighted. The table shows accuracy: 91.94% +/- 0.21% (micro average: 91.94%). Below this, other metrics like precision, recall, f-measure, sensitivity, and specificity are listed. The interface includes tabs for 'Input Data', 'Other Results', and 'Process'. On the right side, there are filters for 'Last Day' and 'Scoring Time (1,000 Rows)', and a status bar indicating '11 s'.

Figure 19: Production Statistics Features.

The screenshot shows the RapidMiner Studio Enterprise interface. The main window title is "ARDSDeployment: Models". The left sidebar shows a "Deployments" section with a "Naive Bayes" entry. The main content area is titled "Model Details" and contains tabs for "Production Model", "Input Data", "Other Results", and "Process". The "Process" tab is selected, showing a detailed XML code representation of the generated process. The XML code is as follows:

```

<?xml version="1.0" encoding="UTF-8"?><process version="9.6.000-RC">
  <context>
    <input/>
    <output/>
    <operator activated="true" class="process" compatibility="9.4.000" expanded="true" name="Process" origin="GENERATED_AUTO_MODEL">
      <parameter key="logverbosity" value="warning"/>
      <parameter key="operator_id" value="17" />
      <parameter key="send_mail" value="never" />
      <parameter key="notification_email" value="" />
      <parameter key="process_duration_for_mail" value="30" />
      <parameter key="env" value="SYSTEM" />
      <process expanded="true" />
      <operator activated="true" class="subprocess" compatibility="9.6.000-RC" expanded="true" height="68" name="Load and Process Data" origin="GENERATED_AUTO_MODEL" width="140">
        <operator activated="true" class="retrieve" compatibility="9.6.000-RC" expanded="true" height="68" name="Retrieve Data" origin="GENERATED_AUTO_MODEL" width="140">
          <parameter key="repository_entry" value="/local Repository/ARDSDeployment/579e78cfa-90cf-4343-9967-8962f7ah5cf4/ARD51" />
          <description align="center" color="transparent" colored="false" width="126">load data.</description>
        </operator>
        <operator activated="true" class="subprocess" compatibility="9.6.000-RC" expanded="true" height="82" name="Create Single Row of Input Data" origin="GENERATED_AUTO_MODEL" width="140">
          <operator activated="true" class="multiply" compatibility="9.6.000-RC" expanded="true" height="103" name="Keep Data for Next Step" origin="GENERATED_AUTO_MODEL" width="140">
            <parameter key="operator_id" value="18" />
            <description align="center" color="transparent" colored="false" width="120">Keep input data for next step.</description>
          </operator>
          <operator activated="true" class="multiply" compatibility="9.6.000-RC" expanded="true" height="124" name="Branch by Type" origin="GENERATED_AUTO_MODEL" width="140">
            <description align="center" color="transparent" colored="false" width="120">Branch processing for categorical and numerical columns.</description>
          </operator>
        </operator>
        <operator activated="true" class="subprocess" compatibility="9.6.000-RC" expanded="true" height="82" name="Handle Dates" origin="GENERATED_AUTO_MODEL" width="140">
          <operator activated="true" class="select_attributes" compatibility="9.6.000-RC" expanded="true" height="82" name="Select Attributes (12)" origin="GENERATED_AUTO_MODEL" width="140">
            <parameter key="attribute_type" value="value_type" />
            <parameter key="use_attributes" value="true" />
            <parameter key="use_except_attributes" value="false" />
            <parameter key="use_except_expression" value="false" />
            <parameter key="value_type" value="date_time" />
            <parameter key="use_value_type_exception" value="false" />
            <parameter key="except_value_type" value="text" />
            <parameter key="use_block_type" value="block" />
            <parameter key="use_block_type_exception" value="false" />
            <parameter key="except_block_type" value="value_matrix_row_start" />
            <parameter key="invert_selection" value="false" />
            <parameter key="include_special_attributes" value="false" />
          </operator>
        </operator>
      </operator>
    </process>
  </operator>
</process>

```

Figure 20: XML Code Process.

The screenshot shows the RapidMiner Studio Enterprise interface. The main window title is "ARDSDeployment: Models". The left sidebar shows a "Deployments" section with a "Naive Bayes" entry. The main content area is titled "Model Details" and contains tabs for "Production Model", "Input Data", "Other Results", and "Process". The "Process" tab is selected, showing a large XML code representation of the generated process. The XML code is as follows:

```

<?xml version="1.0" encoding="UTF-8"?><process version="9.6.000-RC">
  <context>
    <input/>
    <output/>
    <operator activated="true" class="process" compatibility="9.4.000" expanded="true" name="Process" origin="GENERATED_AUTO_MODEL">
      <parameter key="logverbosity" value="warning"/>
      <parameter key="operator_id" value="17" />
      <parameter key="send_mail" value="never" />
      <parameter key="notification_email" value="" />
      <parameter key="process_duration_for_mail" value="30" />
      <parameter key="env" value="SYSTEM" />
      <process expanded="true" />
      <operator activated="true" class="subprocess" compatibility="9.6.000-RC" expanded="true" height="68" name="Load and Process Data" origin="GENERATED_AUTO_MODEL" width="140">
        <operator activated="true" class="retrieve" compatibility="9.6.000-RC" expanded="true" height="68" name="Retrieve Data" origin="GENERATED_AUTO_MODEL" width="140">
          <parameter key="repository_entry" value="/local Repository/ARDSDeployment/579e78cfa-90cf-4343-9967-8962f7ah5cf4/ARD51" />
          <description align="center" color="transparent" colored="false" width="126">load data.</description>
        </operator>
        <operator activated="true" class="subprocess" compatibility="9.6.000-RC" expanded="true" height="82" name="Create Single Row of Input Data" origin="GENERATED_AUTO_MODEL" width="140">
          <operator activated="true" class="multiply" compatibility="9.6.000-RC" expanded="true" height="103" name="Keep Data for Next Step" origin="GENERATED_AUTO_MODEL" width="140">
            <parameter key="operator_id" value="18" />
            <description align="center" color="transparent" colored="false" width="120">Keep input data for next step.</description>
          </operator>
          <operator activated="true" class="multiply" compatibility="9.6.000-RC" expanded="true" height="124" name="Branch by Type" origin="GENERATED_AUTO_MODEL" width="140">
            <description align="center" color="transparent" colored="false" width="120">Branch processing for categorical and numerical columns.</description>
          </operator>
        </operator>
      </operator>
    </process>
  </operator>
</process>

```

Figure 21: GLM with Feature Generation Model.

The screenshot shows the RapidMiner Studio interface with the 'Auto Model' tab selected. The process flow is shown as a horizontal bar with six steps: Load Data, Select Task, Prepare Target, Select Inputs, Model Types, and Results. The 'Results' step is highlighted with an orange circle. Below the process bar are buttons for RESTART, BACK, OPEN PROCESS, EXPORT, and DEPLOY. On the left, a sidebar titled 'Results' shows a tree view with 'Generalized Linear Model' expanded, listing Model, Simulator, Performance, Lift Chart, Feature Sets, Predictions, and Production Model. Under 'General', there is a 'SAVE RESULTS' button. The main content area displays the 'Generalized Linear Model - Model' results table:

| Attribute | Coefficient | Std. Coefficient | Std. Error | z-Value | p-Value |
|-----------|-------------|------------------|------------|---------|---------|
| bmi | -0.003 | -0.028 | ? | ? | ? |
| pf | 0.062 | 7.937 | ? | ? | ? |
| log(pf) | 29.883 | 6.384 | ? | ? | ? |
| Intercept | -92.762 | -8.326 | ? | ? | ? |

Figure 22: GLM with Feature Generation Production Model.

The screenshot shows the RapidMiner Studio interface with the 'Auto Model' tab selected. The process flow is identical to Figure 21. The 'Results' step is highlighted with an orange circle. The sidebar on the left shows the 'Production Model' node selected under 'Feature Sets'. The main content area displays the 'Generalized Linear Model - Production Model' results table:

| Attribute | Coefficient | Std. Coefficient | Std. Error | z-Value | p-Value |
|-----------|-------------|------------------|------------|---------|---------|
| bmi | 0 | 0 | ? | ? | ? |
| pf | 0.062 | 7.952 | ? | ? | ? |
| log(pf) | 30.061 | 6.419 | ? | ? | ? |
| Intercept | -93.293 | -8.345 | ? | ? | ? |

Figure 23: Simulator GLM Model with Feature Generation.

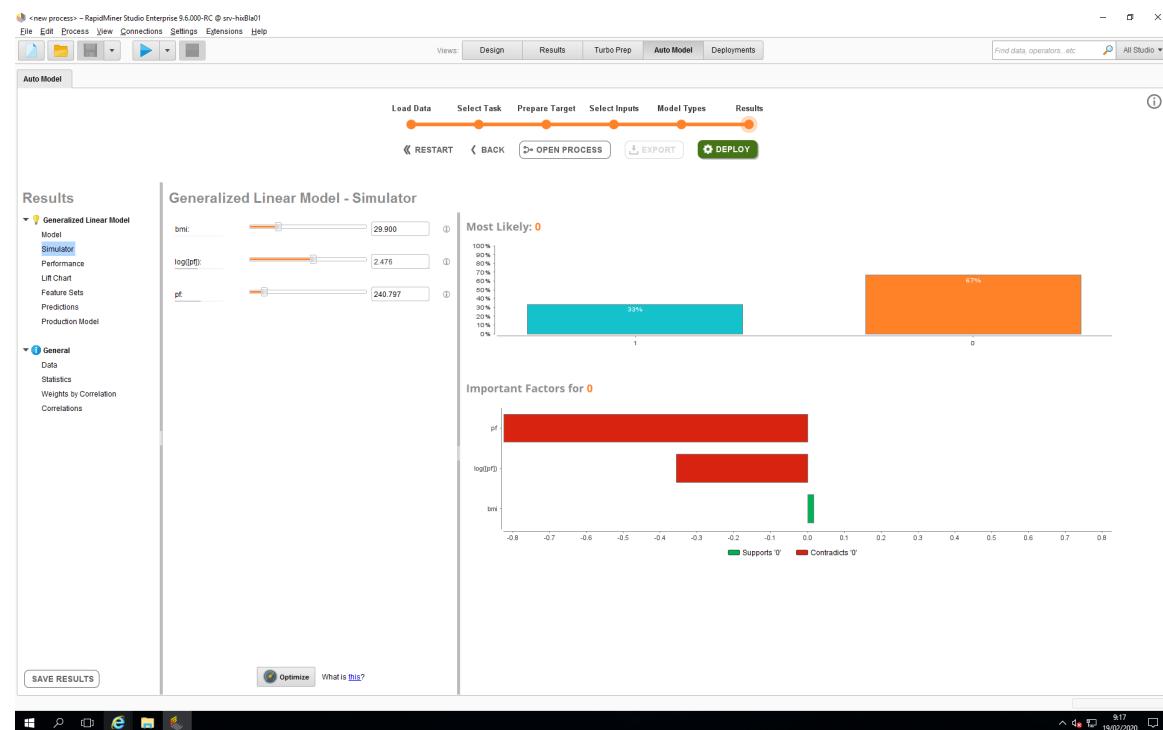
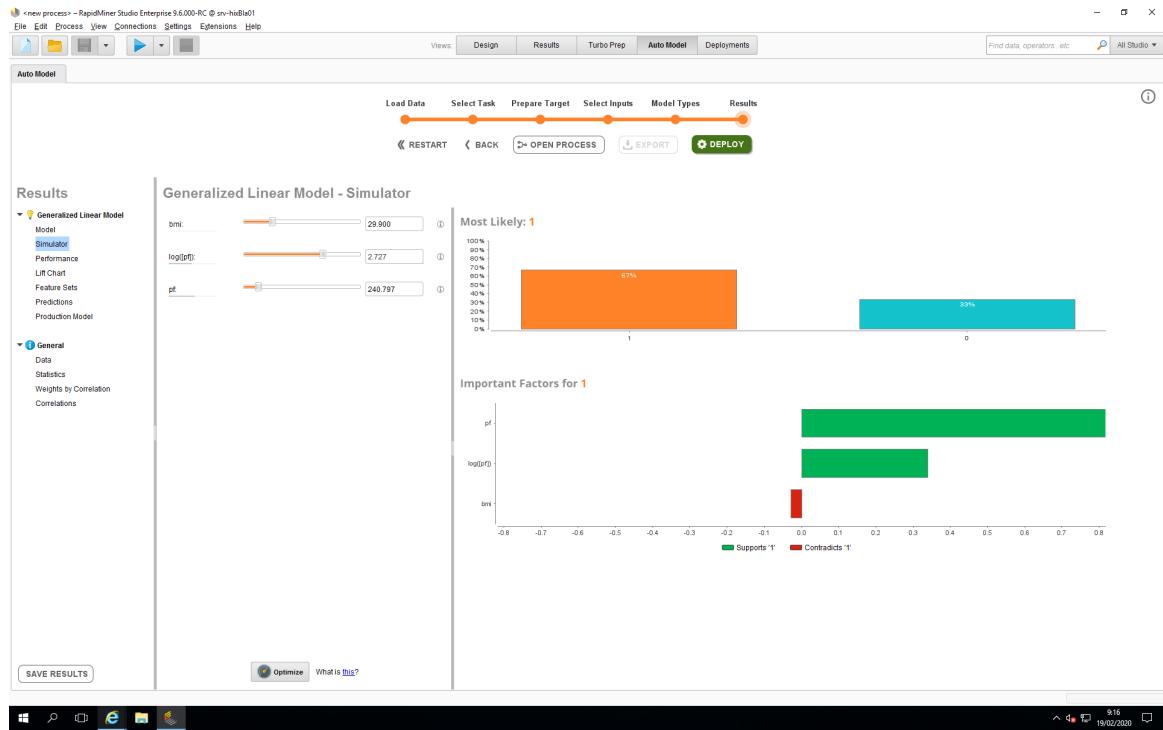


Figure 24: GLM with Feature Generation Performance.

The screenshot shows the RapidMiner Studio Enterprise interface for a 'Generalized Linear Model - Performance' process. The process flow consists of six steps: Load Data, Select Task, Prepare Target, Select Inputs, Model Types, and Results. Buttons for RESTART, BACK, OPEN PROCESS, EXPORT, and DEPLOY are located below the flow. The 'Results' pane on the left is expanded to show the 'Performance' tab, which displays various metrics:

| Criterion | Value | Standard Deviation |
|----------------------|--------|--------------------|
| Accuracy | 99.7% | ± 0.0% |
| Classification Error | 0.3% | ± 0.0% |
| AUC | 100.0% | ± 0.0% |
| Precision | 100.0% | ± 0.0% |
| Recall | 98.6% | ± 0.1% |
| F Measure | 99.3% | ± 0.1% |
| Sensitivity | 98.6% | ± 0.1% |
| Specificity | 100.0% | ± 0.0% |

Figure 25: GLM Feature Generation, Feature Set.

The screenshot shows the RapidMiner Studio Enterprise interface for a 'Generalized Linear Model - Feature Sets' process. The process flow is identical to Figure 24. The 'Results' pane is expanded to show the 'Feature Sets' tab, which includes a scatter plot titled 'Optimal Trade-offs between Complexity and Error'. The plot shows the relationship between Complexity (Y-axis, 0 to 40) and Error (X-axis, 0.00% to 6.00%). A legend indicates: Used feature set (blue dot), Optimal trade-offs (green dots), Original feature set (grey square), and Shown below (orange dot). Below the plot, a description explains the purpose of the feature engineering run and the meaning of the different data points. The 'Currently Selected Feature Set' table lists the following features:

| Name | Expression | Complexity |
|----------|------------|------------|
| pf | [pf] | 1 |
| bmi | [bmi] | 1 |
| GenSym70 | log(pf) | 2 |

Figure 26: Correlations based on GLM Model.

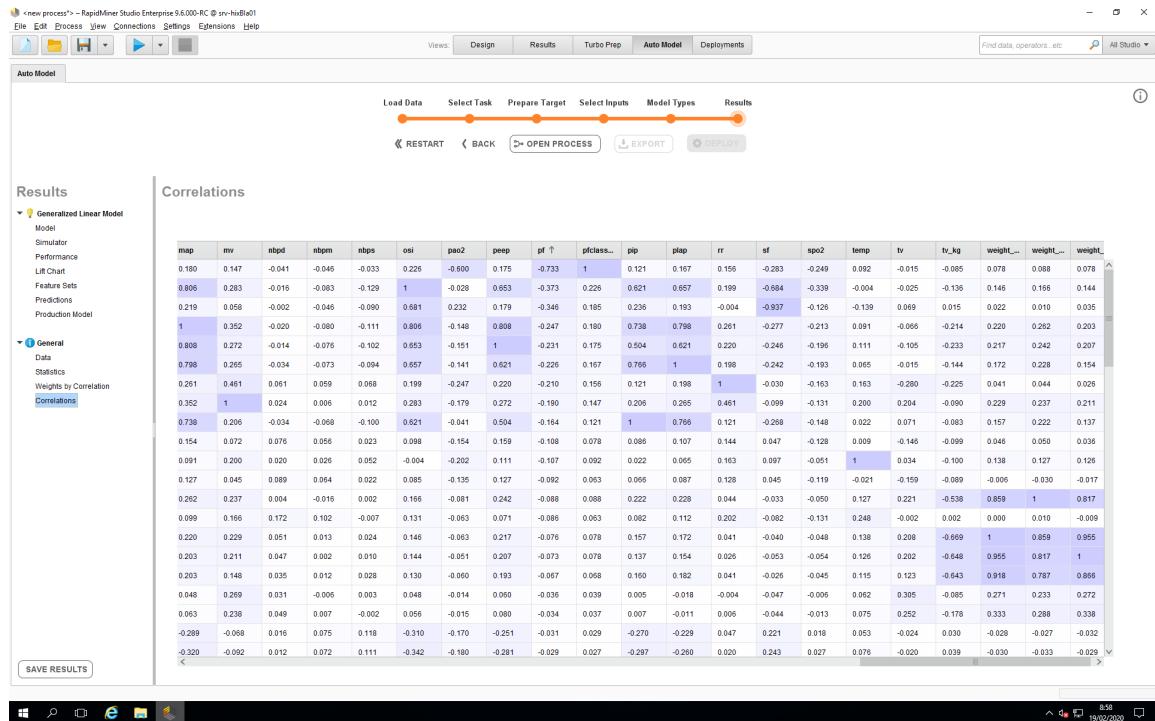
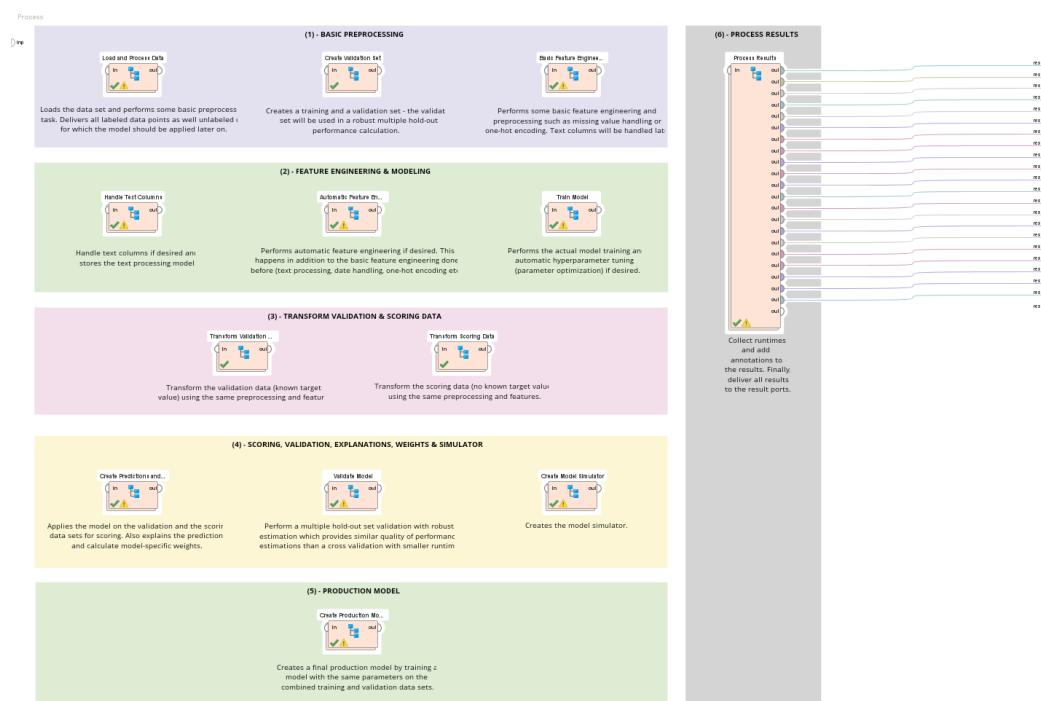


Figure 27: Visual Representation of the Process from Preprocessing to Production Model.



Discussion

The approach presented in this document never intended questioning the results of the paper as published by Pengcheng Yang et al. However the technology available from RapidMiner enabling automated predictive modeling, model deployment and monitoring are illustrated. In particular the presented tools should stimulate a feasible implementation in any medical setting or environment. The most relevant arguments for a more widespread use of this approach are:

1. Code free deployment
2. Deploy models without the need for technical skills
3. Manage multiple deployment locations and share deployed models with other users
4. Drive collaboration on models, monitor for governance, drift and bias issues and even set up alerts and integrations
5. See how models work in the wild and adjust your strategy accordingly
6. Deploy multiple models together where one model is active and others are ‘challengers’, where predictions and error rates are also stored for challengers
7. Review performance of active and challenger models over time on the leaderboard
8. Setup alerts to notify you when a challenger outperforms the active model
9. Swap the active model with a single click so you’ve always got the best model deployed into your workflow
10. Go beyond ‘accuracy’ stats to understand and demonstrate medical outcome and or financial impact of your models
11. Show cumulative gains and impact produced by your active model
12. Analyze scoring times to pick models fast enough for your needs
13. Compare distribution differences between predictions and actual values
14. Instantly understand problematic trends and address them proactively
15. Understand models and see how they behave with weights, model visualizations, interactive simulators and full prediction explanations
16. Compare models’ latest performance with expected error rates to detect concept drift or shift
17. Calculate input factor drift and compare drift with factor importance as early indicators for problematic concept shift or drift
18. Observe suspicious changes between training and scoring data and detect bias in training data