Price Prediction in Railway Tenders using Machine Learning

Pablo Garrido Martínez–Llop
Department of MechE, UPM
Madrid, Spain
Juan de Dios Sanz Bobi
Department of MechE, UPM
Madrid, Spain
Carlos Simancas Cuevas
Department of MechE, UPM
Madrid, Spain

ABSTRACT
Over the time, the number of public or private tenders among the competitiveness increases considerably, so companies must choose carefully which ones to bid for. For this selection two main key aspects should be identified: tenders in which the company has more competitive advantage and the maximum profit margin that the company can assume without the risk of not being awarded due to better economical offers of other competitors. In this paper, a decision tool to help railway companies to make decisions regarding the participation in certain bidding processes is developed applying Machine Learning algorithms. A complete database of 480 data samples including 17 different variables in each sample is used for training the four algorithms: ridge regression, random forest, support vector machine (SVM) and Artificial Neural Networks (ANN). The variable to be predicted is the competitors’ bid price range for commuter trains in the railway sector. A comparative table with the performance of each model is presented, concluding the most appropriate one to be used depending on the different price ranges established.

General Terms
pricing, price prediction, bidding process optimization, machine learning real application, industry 4.0, data automation and data mining.

Keywords
pricing, price prediction, machine learning, ridge regression, random forest, support vector machine (SVM), Artificial Neural Networks (ANN).

1. INTRODUCTION
A tender is a regulated process by which a private or public organisation, publicly requests bids to cover a service or necessity. In the railway context, the railway operating companies establish a technical specification with all the requirements and characteristics such as the number of trains that are needed, the number of coaches in each train, the maximum operating speed, the track gauge,... After all bidding offers are received, a detailed evaluation is developed to determine the awarded company. Normally, there are two main parts to evaluate: the technical and the economical part.

The commitment to rail as a safe and environmentally efficient mean of transport is leading to the growth of new rail tenders in all Central Europe [1]. On the other hand, competitiveness among companies is also increasing exponentially. For these reasons, an appropriate selection of the tenders worth betting on becomes an essential key for the midterm strategy of railway manufacturers.

A key element for the above-mentioned purpose is to know an estimated price range of the competitors and therefore, the most likely tenders to be successful in the award process.

Predicting the most competitive price with high accuracy and precision is very difficult because not only objective data are involved, but also depends on other subjective aspects such as the convenience for bidding up or down due to strategic decisions or even the exact margin that a manager wishes to obtain in each project [2].

Based on different pricing strategies of companies, trying to identify the key elements in decision-making and which techniques are most appropriate depending on the sector, the company's situation and the product and market context, a technological tool is created [3]. The main target of this tool is to predict the price range in which the most competitive price for a particular railway tender can be found, given the technical characteristics and context of interest [4]. In other markets demand forecasting and search for optimal prices is being carefully analyse so the extrapolation to the railway sector should be investigated [5].

For the prediction model definition, machine learning techniques have been applied [6], starting from a historical database of 480 previous tenders, containing technical data and the price of the final awarded contract. Four different supervised learning algorithms, including Artificial Neural Networks [7], are compared and tested, concluding the most appropriate one for each price range and train characteristics.

2. THEORETICAL BASE
The four algorithms used in this paper are: ridge regression, random forest, support vector machine (SVM) and artificial neural networks (ANN).

This selection is based on the different characteristics of the problem to be solved: a dataset containing between 50 and 1000.000 samples with more than 10 influential variables involved with no key variables detected and with an acceptable output of an approximate range of prices (it is never expected an exact value because subjective aspects are involved as the decision of the gross margin to be included).

Ridge regression is based on linear regression considering a set of variables. The main expression is: \( y = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n \), where:

- \( y \) represents the output of the model or label,
- \( w = (w_0, w_1, w_2, \ldots, w_n) \) is a vector containing the weights to be optimized by the model,
- \( x = (x_0, x_1, x_2, \ldots, x_n) \) is the input vector with the characteristics of the model.

The main goal is to locate the hyperplane that best fits the multidimensional data. It is useful for cases in which a simple linear approach is enough for the purpose given, not being able to find non-linear patterns in the data. The performance of ridge regression is evaluated using the quadratic error to obtain the optimal values of the weights (w). These weights

\[ y = \begin{bmatrix} x_1 \cdots x_n \end{bmatrix} \begin{bmatrix} w_0 \cdots w_n \end{bmatrix} \]
are updated in the different iterations of the model, so the total sum of errors is minimized. In this algorithm a new correction term is added to avoid model overfitting.

An Artificial Neural Network (ANN) is a computer algorithm based on a set of simple, constructively identical nodes, called artificial neurons or perceptrons. These perceptrons are grouped in layers, so that all the perceptrons in one layer are connected to all the perceptrons in the next layer, allowing the transmission of information between layers of ANN.

In analogy to biological neurons, an artificial neuron is defined as a computational node that receives the values of the activations computed by the neurons of the previous layer as input. At the same time, computes the activation that will be sent to downstream neurons. Each neuron is characterized by the following parameters: \( w_i \) (weight) and \( b \) (independent term). The structure of an artificial neuron could be appreciated in Figure 1.

![Figure 1. Structure of an artificial neuron](image)

The inputs or activations received are subjected firstly to a linear operation, and secondly to a non-linear operation. In linear operation, the independent weights and terms learned by the model are used according to the expression:

\[
z = \sum_i (w_i \cdot a_i) + b
\]

The non-linear operation or activation phase is implemented by ReLU and sigmoid functions. In turn, each neuron could be contained within three different types of layers: input layer, hidden layers and output layer. The number of hidden layers in a neural network depends on the problem to be solved and is a hyper parameter of the model that must be optimized to maximize the accuracy of the model. A high number of layers implies a deeper and more complex neural network. The optimal number of layers should be determined.

Random Forest uses a random distribution of a predetermined number of decision trees, considering the average of all the results as the final value to be implemented. A decision tree with \( p \) input variables works as follows:

- The \( p \) – dimensional feature space is divided into \( M \) mutually distinct regions that cover entirely the subset of the feature space and do not overlap.
- The value range of each variable is divided in two parts (nodes). The partition that minimizes the total error is selected and the process continues creating new nodes. This iterative algorithm ends once an acceptable error is determined.
- For the prediction process a new case is pushed down the tree and the label of the terminal node is assigned.

Support Vector Machine (SVM) is an algorithm used for category classification so a hyperplane that optimally separates two existing classes is defined. Data may have been previously projected to a higher dimensional space.

The hyperplane to be defined should have the maximum distance (margin) to the points that are closest to itself. The vector formed by the points closest to the hyperplane is called the support vector. With this maximum margin goal, the hyperplane capable of classifying (or separating) new data is obtained.

For regression problems, decision boundary hyperplanes are defined, and the main goal is to achieve the hyperplane that contains the maximum number of samples.

3. METHODOLOGY

After selecting the algorithms to be tested, the following steps are necessary to solve the problem:

3.1 Prepare the datasets.

A database containing 480 has been prepared in this paper containing 17 different variables in each sample. This information is obtained from historical data on public tenders resolved in the past.

The 17 variables are: the country in which the tender is published, the track gauge, the number of floors of the train (1 or 2), the maximum speed of the train, the bid price of the awarded company (variable to be predicted), the number of trains required, the number of coaches per train, the passenger’s capacity of each train, the train total length, the number of axles of the train, the number of voltages required in the train for traction purpose, the place for manufacturing the train, the price of the steel, the price of the aluminium, the GDP per capita in the country in which the tender is published, the year of publication of the tender and the total number of coaches.

Not all the variables are equally important in price prediction, so mathematic analysis such as the PCA (Principal Component Analysis) is implemented in order to separate the most important variables from the rest.

Different databases are created for obtaining the final best database for running the defined models:

Database 1 (DB1): first of all, the number of coaches per train is removed as it could be considered redundant information as the total number of coaches and the number of trains are also collected in the database. A total number of variables of 16 is considered.

Database 2 (DB2): A first PCA is applied to DB1 so a total number of variables of 13 is considered. The following variables are discarded:

- The country: as the manufacturing companies are international expanded companies, the manufacturing process is fixed regardless the country in which the tender is published.
- The track gauge: different track gauge has no additional economic implication in the design and manufacturing process of a train.
- The place for manufacturing the train: can be considered redundant since the GDP per capita in the country in which the tender is published is also included in the database.

Database 3 (DB3): A second PCA is applied so a total number of variables of 11 is considered. The following additional variables are discarded:

- The train length: it is directly related with the train capacity considering also that a maximum length is permitted by regulations.
The number of voltages required in the train for traction purpose; this technical requirement affects in materials and design complexity, but it is not very relevant compared to the total cost of the train.

Database 4 (DB4): A third PCA is applied so a total number of variables of 10 is considered. The following additional variables are discarded:

- The number of axles of the train: this variable could be related with the number of coaches and the total passenger’s capacity of the train.

Database 5 (DB5): this database is obtained selecting the 11 most relevant variables from the point of view of a group of experts belonging to a rolling stock manufacturer considering the process done in the four previous databases created. The 11 variables are: the maximum speed of the train, the bid price of the awarded company (variable to be predicted), the number of trains required, the number of coaches per train, the number of axles of the train, the number of voltages required in the train for traction purpose, the price of the aluminium, the GDP per capita in the country in which the tender is published, the year of publication of the public tender and the total number of coaches.

### 3.2 Data processing

Reliability and integrity of data is essential for a good modelling and results. In fact, data processing is the key for creating successful models using different algorithms. Databases used in this paper has been completed, reviewed and processed. Some data were missing and specific suppression of outliers in the model is done. Not reliable or inconsistent samples have also been removed and some characteristics are added such as the aluminium or steel price after expert consultation.

### 3.3 Model construction and implementation

All variables expressed in string format had to be converted to numerical variables, creating as many binary variables as necessary. The next step is the selection of the predictable variable ("y"), which is the bid price of the awarded company. This is the variable that we would like to predict with our final model. The rest of the values are collected in the matrix “X” and all of them are previously normalized into a 0 to 1 range. Data normalization is relevant to make model training less sensitive to the scale of characteristics involved in the database.

### 3.4 Running the models

In order to evaluate the models, two different sets are defined: the training set, used for training the models; and the test set, used for validating the models. The training set contains the 90% of the samples and the test set contains the 10% of the samples. Samples are randomly selected for each set and models are trained and evaluated 10 times with 10 different data distribution between both sets to assure the correct performance evaluation of the model. The final evaluation is done considering the average of the 10 evaluations done.

The models that have been run in this paper are evaluated considering the coefficient of determination, R2. This coefficient represents the amount of variation of the predicted price of the awarded company compared with the real value. A negative or 0 value of this coefficient determines that the model’s performance is not acceptable. If $R^2 = 1$ the error committed is null and the predictions and the real values are the same. Therefore, the main target is to get as close as possible to $R^2 = 1$ and use this coefficient to discard models and compare models to each other.

### 3.5 Improvement of the useful models

#### Search for the best hyperparameters

In this phase an optimization of the model is reached adjusting different hyperparameters, which are configurable model values that could be modified by the user to check the most appropriate ones for the best performance of the model. These optimal values of the hyperparameters are not known in advance for a given problem, so generic recommended values are implemented by default and different tests with different values are developed for being selected by trial and error.

#### 3.6 Results and performance of the models

In this last phase, all results and performance of the created models are exposed and compared in order to select the most appropriate model for each case, as different algorithms with different databases are tested.

### 4. MACHINE LEARNING SUPERVISED MODELS AND RESULTS

For the evaluation of each model, the different hyperparameters of the algorithms used are determined first:

- **Ridge Regression**: the hyperparameter $\alpha$, which indicates the level of regularization of the parameters, has been changed from default $\alpha=1$ to $\alpha=0.5$ so a better performance is obtained.

- **SVM**: linear kernel is established for a better result.

- **Random Forest**: the number of estimators or trees that reaches the optimal results are obtained with 10 instead of the default value of 100.

- **ANN**: based on experience with similar studies a ten-layer network with fifty neurons in each layer is implemented. The regularization hyperparameter considered is 0.001 and the learning rate is stabilized in 0.5-10.

The first decision to be made is related with the final database to be used in the different models. In this sense, a first running of the models is done to select the most appropriate database comparing the R2 results. The results are presented in Figure 2.

<table>
<thead>
<tr>
<th></th>
<th>DB1</th>
<th>DB2</th>
<th>DB3</th>
<th>DB4</th>
<th>DB5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ridge Reg.</td>
<td>0.8065</td>
<td>0.8606</td>
<td>0.8624</td>
<td>0.865</td>
<td>0.8730</td>
</tr>
<tr>
<td>SVM</td>
<td>0.7917</td>
<td>0.8087</td>
<td>0.8062</td>
<td>0.8056</td>
<td>0.8216</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.7512</td>
<td>0.7692</td>
<td>0.7319</td>
<td>0.7702</td>
<td>0.8054</td>
</tr>
<tr>
<td>ANN</td>
<td>0.65</td>
<td>0.7271</td>
<td>0.6973</td>
<td>0.7143</td>
<td>0.8313</td>
</tr>
<tr>
<td>Average</td>
<td>0.7499</td>
<td>0.7914</td>
<td>0.7745</td>
<td>0.7888</td>
<td>0.8409</td>
</tr>
</tbody>
</table>

Figure 2. R2 values obtained running the models with the different databases created.

It is concluded that the optimal database to be used for the models is DB5, which is a mixture of expert consultation also considering statistics studies (PCA) of the previous databases.

For a correct evaluation of each model, as it is stated in the previous section, training and test sets are defined respecting a 90:10 rate and 10 error measures are determined due to the 10
different evaluations done for each model defined. In total 40 evaluations are carried out: 10 different evaluations of each of the 4 algorithms implemented.

Depending on the contract price ranges different algorithms could offer the best accuracy and precision in the results. For this reason, after all evaluations the results are divided into 4 ranges in order to check the most appropriate model for each range. This division is done equilibrating also the number of samples in each range as it is appreciated in Figure 3:

<table>
<thead>
<tr>
<th>Contract Price Ranges (M€)</th>
<th>Number of samples in each range</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-50</td>
<td>150</td>
</tr>
<tr>
<td>50-100</td>
<td>121</td>
</tr>
<tr>
<td>100-200</td>
<td>96</td>
</tr>
<tr>
<td>&gt;200</td>
<td>113</td>
</tr>
</tbody>
</table>

Figure 3. Ranges established after running the four different algorithms

After running the 40 different evaluations and comparing the average of the relative errors obtained in each of the 10 evaluations made for each algorithm it is possible to summarize the final results, classified in different ranges. These results are presented in Figure 4.

<table>
<thead>
<tr>
<th>Contract Price (M€)</th>
<th>Ridge Regression</th>
<th>SVM</th>
<th>Random Forest</th>
<th>Red Neuronal</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-50</td>
<td>65.15%</td>
<td>60.99%</td>
<td>26.36%</td>
<td>60.37%</td>
</tr>
<tr>
<td>50-100</td>
<td>30.39%</td>
<td>20.86%</td>
<td>22.78%</td>
<td>27.04%</td>
</tr>
<tr>
<td>100-200</td>
<td>18.78%</td>
<td>13.34%</td>
<td>24.21%</td>
<td>17.15%</td>
</tr>
<tr>
<td>&gt;200</td>
<td>19.80%</td>
<td>25.37%</td>
<td>22.52%</td>
<td>22.15%</td>
</tr>
</tbody>
</table>

Figure 4. Final results of the relative error of the models classified in the different ranges

For samples corresponding to contract awards under 50 M€, Random Forest is the most appropriate algorithm to predict this value. If the contract award is between 50 and 200 M€, then the best option is to test the SVM algorithm. For larger contracts Ridge Regression should be used. Relative error’s average is near 20% which means a good first approximation of the competitor’s economical offers.

Finally, the influence of the variables selected in the final results is measured to validate the expert criteria. The result is shown below in Figure 5:

The most influential variables are: the total number of coaches, the aluminium price, the number of trains, the year of the contract award and the GDP per capita in the country where the tender is published.

On the other hand, the number of axles, the maximum speed, the number of floors and voltages of the train and the number of coaches per train are less important. All of these variables could be redundant except the maximum speed, which all experts expected to be a very relevant variable for the final price of the contract.

5. CONCLUSIONS

A reliable and robust database should include rational statistic studies with technical experts’ validation, as the variables proposed in the database 5 (DB5) led to better results running the algorithms.

The most relevant variable for the price range prediction of a railway contract is the total number of coaches required, so a good strategy for railway companies is to optimize in the offers the number of coaches fulfilling all the rest of requirements: the number of seats, the distances between toilets, moreover, the maximum speed to be reached by the train is not so relevant as expected by the experts.

The algorithms’ performance is not the same for all the price ranges established, so the correct approach is to differentiate the best algorithm to be used in each range to predict the outcome of the bidding process. In Figure 6 the best performing algorithms are presented depending on the set range and with the minimum error that has been achieved.

<table>
<thead>
<tr>
<th>Contract Price (M€)</th>
<th>Best Algorithm</th>
<th>Error Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-50</td>
<td>RANDOM FOREST</td>
<td>26.36%</td>
</tr>
<tr>
<td>50-100</td>
<td>SVM</td>
<td>20.86%</td>
</tr>
<tr>
<td>100-200</td>
<td>SVM</td>
<td>13.34%</td>
</tr>
<tr>
<td>&gt;200</td>
<td>RIDGE REGRESSION</td>
<td>19.80%</td>
</tr>
</tbody>
</table>

Figure 6. Performance of the algorithms depending on the price range

Random Forest algorithm is really useful for small contracts up to 50 M€, as it is stated in Figure 6 compared with the rest of the algorithms.

As a principal conclusion, a decision tool to help railway companies to make decisions regarding the participation in certain bidding processes has been developed. This tool could allow these companies to centre the strategy and concentrate in the most appropriate offers where the company is more competitive. The range of the contract prices could be predicted with acceptable accuracy for the purpose, considering relative errors around 20%.

Lastly, it should be mentioned that performance could be improved, and relative error reduced with the acquisition of much more data to work with. This improvement could not only allow to have a decision maker tool, but also to precise the maximum profit margin that certain company could include assuring the contract award.
6. ACKNOWLEDGMENTS
The authors acknowledges all rolling stock experts that have contributed in obtaining the input necessary information for this machine learning applied case.

7. REFERENCES


