HiTEc meeting &
Workshop on Complex data in
Econometrics and Statistics

# Machine Learning and Uncertainty Analysis for Remaining Value Estimation

**Ieva Dundulienė (KTU), Robertas Alzbutas (KTU, LEI)**

HiTEc & CoDES, 2024/03/24

European Cooperation in Science & Technology
Funded by the European Union

# Outline

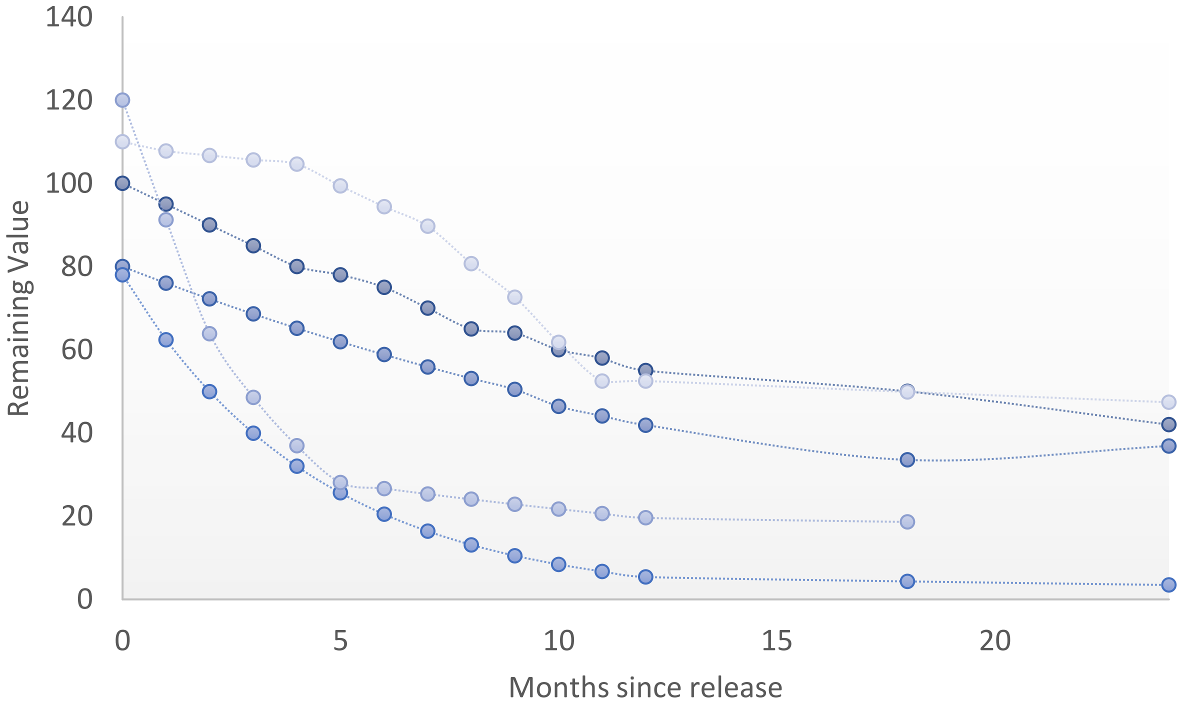
- Remaining Value Estimation
- Uncertainty Analysis, Problems & Opportunities
- Gold Standard in Refurbished Products Evaluation
- Machine Learning Application for Refurbished Products Remaining Value Estimation
- Case Study: Automated Refurbished Smartphones Remaining Value Estimation

# Remaining Value (RV)

***Remaining Value (RV)*** –

Estimated value of the product at the specific time moment *t*.

Remaining Value depends on multiple factors that should be considered while evaluating RV.



RV variation example of hypothetical product

# Machine Learning (ML) Models Application for Remaining Value Estimation

ML models are commonly used for remaining value estimation for solving problems:

- Predictive/preventive manufacturing line maintenance, reducing downtime and optimizing maintenance schedules.
  *Sensor data analysis, NN, Gradient Boosting, Survival Analysis, Time Series Forecasting*

- Estimating RV or future performance of financial assets, portfolios.
  *Time Series Forecasting (ARIMA, GARCH, …), ML models, Monte Carlo simulations*

- Inventory Management and Supply Chain while forecasting the remaining inventory levels, demand forecasting, managing supply chain operations.
  *Time Series Forecasting, ML models, Stochastic modelling.*

# Problems & Opportunities

## Economic

Recent studies have noted a shift n the world economic model towards ***sustainability*** through the circular economy with ***artificial intelligence*** helping in facilitating this transformation.

## Environmental

***53.6* Mt** generated E-waste in 2019. Every year global e-waste generation is increasing.

.

## Social

Customers are more frequently considering to buy refurbished or used items.

Customers overlook refurbished products due to a lack of awareness of the refurbishment principle.

## Behavioural

Mobile phone penetration is estimated at 90% among EU adults.

# Gold Standard in Refurbished Products Evaluation

- Evaluating the Remaining value of refurbished products creates new challenges due to the refurbished products‘ technical characteristics, unobserved usage patterns, and others.
- The absence of validation standards (etalon) or publicly available standardized data complicates the evaluation process.
- Limited data availability bring additional obstacles while evaluating products characteristics.
- The growing consumer preference for refurbished products in the long term perspective will increase the demand for automated solutions that evaluation.

# Regression Evaluation

If $\bar{y}$ is the mean of observed data

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

The residuals squares calculation

$$SS_{res} = \sum_{i}(y_i - \hat{y}_i)^2 = \sum_{i} {e_i}^2$$

The total sum of squares

$$SS_{tot} = \sum_{i}(y_i - \bar{y})^2$$

**Coefficient of determination**

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

**Other** regression evaluation metrics:

Mean Absolute Error (MAE)

$$MAE = \frac{1}{n}\sum_{i} |y - \bar{y}|$$

Mean Squared Error (MSE)

$$MSE = \frac{1}{n}\sum_{i}(y - \hat{y})^2$$

Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{MSE}$$

# Semi-Supervised learning

**Semi-supervised learning** is a **subset** of Machine Learning (ML) that combines supervised and unsupervised learning practices.

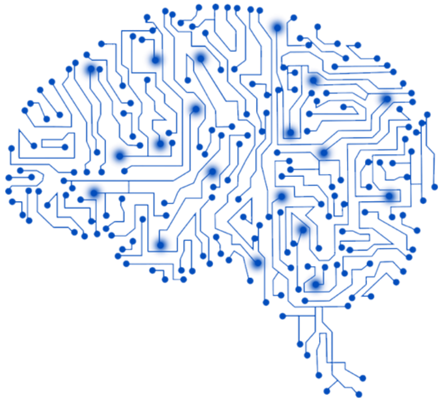
In terms of semi-supervised learning there **might be applied** supervised or unsupervised ML techniques but due to application limitations, results accuracy, and cost semi-supervised learning is usually applied to compensate for these limitations.

**Supervised learning limitations** when there is a limited amount of labeled data:

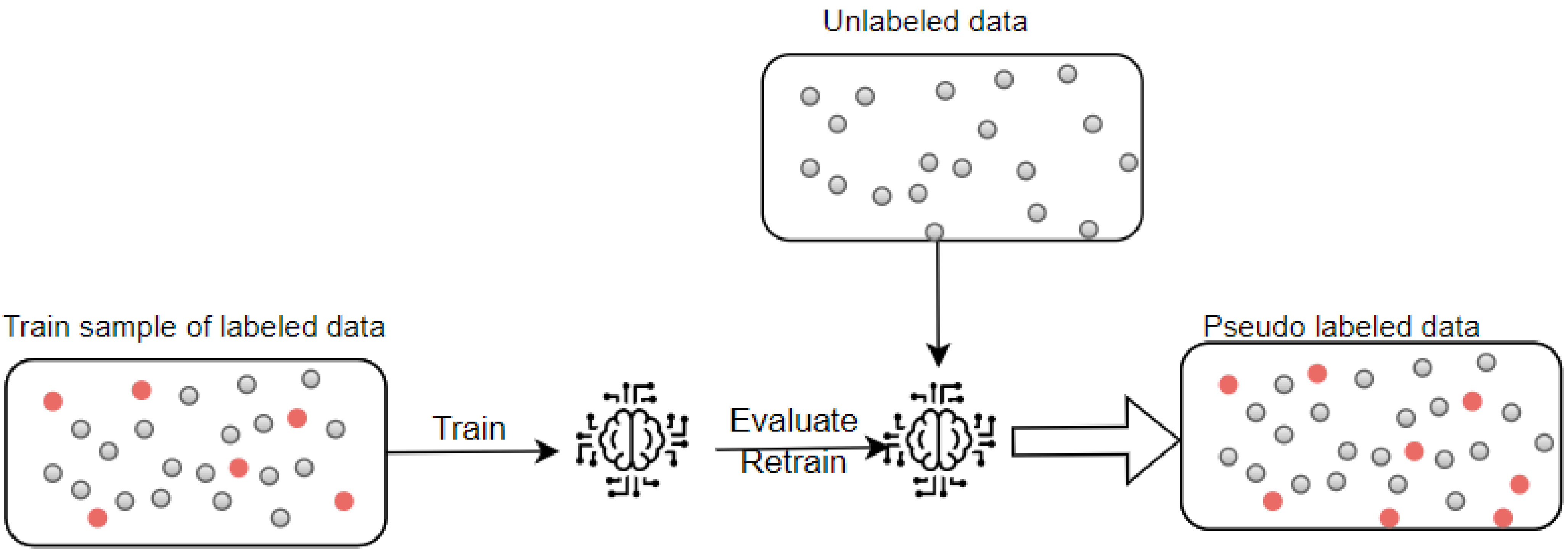
- Slow (Should find an expert to label/validate values)
- Costly (Requires to have a large data volume
  to have accurate predictions)

**Unsupervised learning limitations:**

- Usually used for clustering
- Results are less accurate or difficult to evaluate



# Semi-Supervised learning



- The labelled data sample is trained with a regression model. Model performance evaluated on the labeled validation set.
- Unlabeled dataset is trained by created and validated model to generate pseudo labels.

# Case Study: Automated Refurbished Smartphones Remaining Value Estimation

**Data:**

- Smartphones' test results values
- Unstructured experts' comments
- Data input (even sources) mistakes

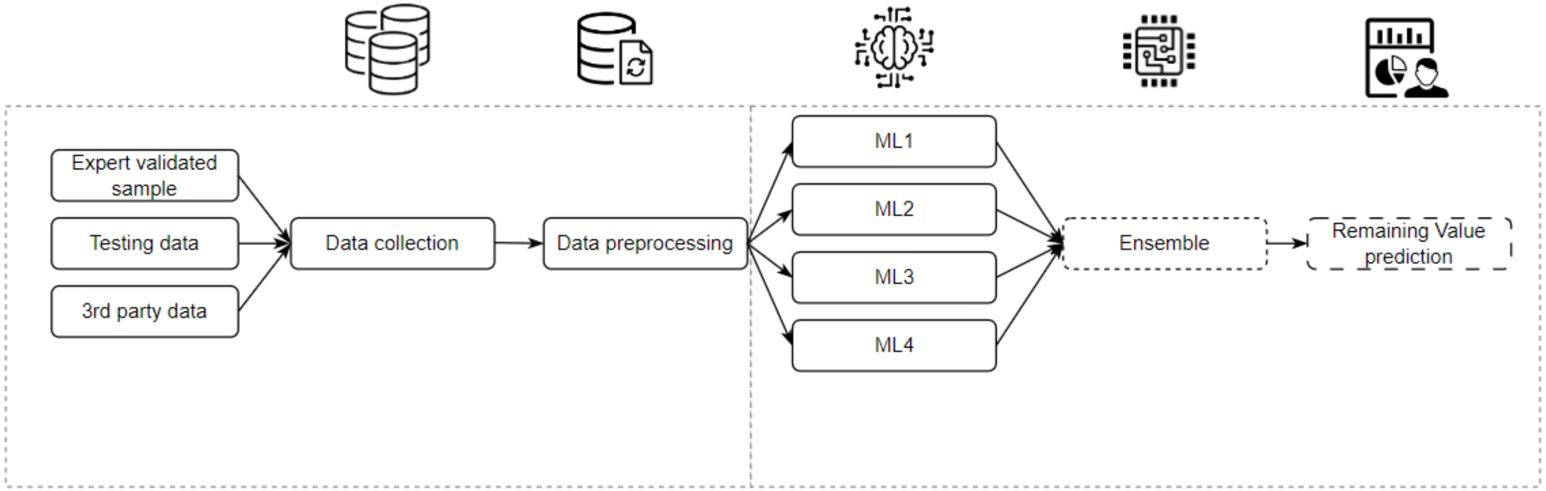
**Key challenges:**

- Multiple data sources and types integration.
- The proportion of known RV is very low (**<10%**)

**Advantages**:

- Structured and automated way of remaining value estimation/comparison.
- Applying Machine Learning Methods and estimating remaining value without known true value.
- Evaluate predicted value uncertainty

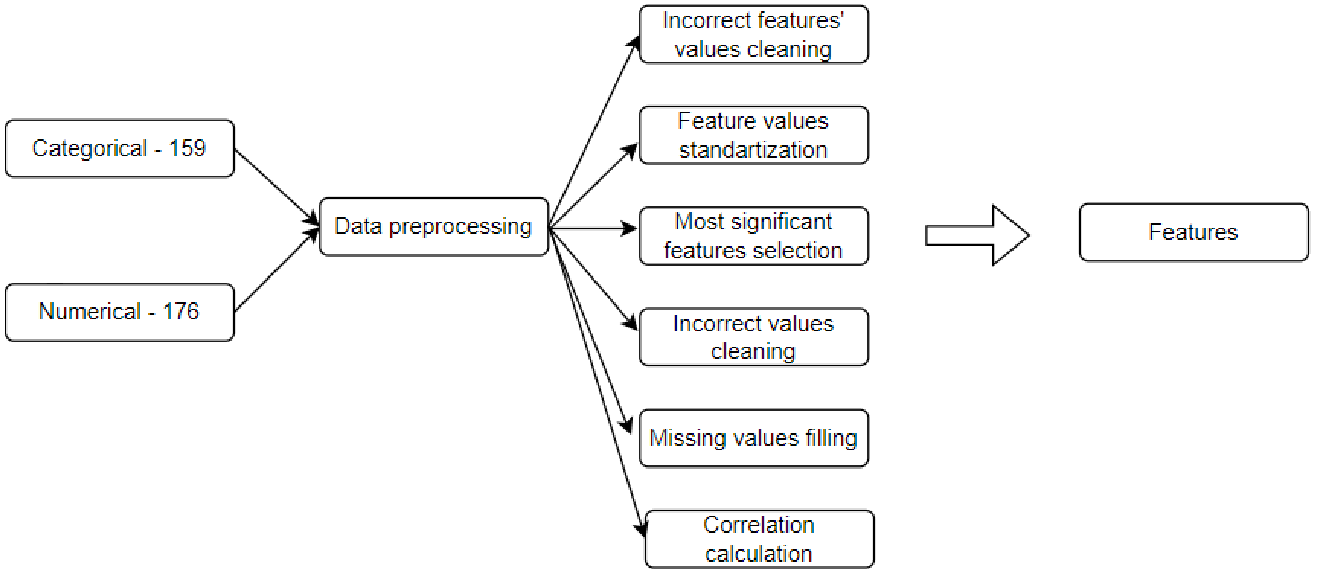
# Case Study: Automated Refurbished Smartphones Remaining Value Estimation Framework



- Semi-supervised learning
- Domain knowledge integration
- Multicollinearity eliminations
- Supervised regression analysis
- Meta model application for Remaining Value (RV) estimation
- Uncertainty Evaluation

# Case Study: Data Preprocessing

- Data preprocessing strategy consists of multiple steps to reduce the number of features, unify feature values, and clean comments.


Categorical - 159
Numerical - 176
Data preprocessing
Incorrect features' values cleaning
Feature values standartization
Most significant features selection
Incorrect values cleaning
Missing values filling
Correlation calculation
Features

- ➢ Unified feature values
- ➢ Written comments validation and preparation
- ➢ Multicollinearity detection
- ➢ Categorical features encoding

- After data preprocessing procedures the final dataset consists of numerical variables (categorical encoded features included)

# Case Study: Multicollinearity

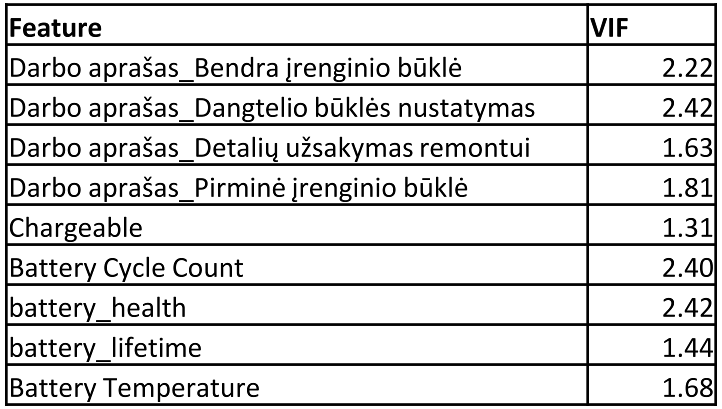
Multicollinearity was identified by multiple methods:

- Pairwise Correlations to estimate pairwise correlations between independent variables.

- Variance Inflation Factor (VIF) evaluates the relationship between one independent variable with all the other independent variables.

*Features excluded when VIF > 10*

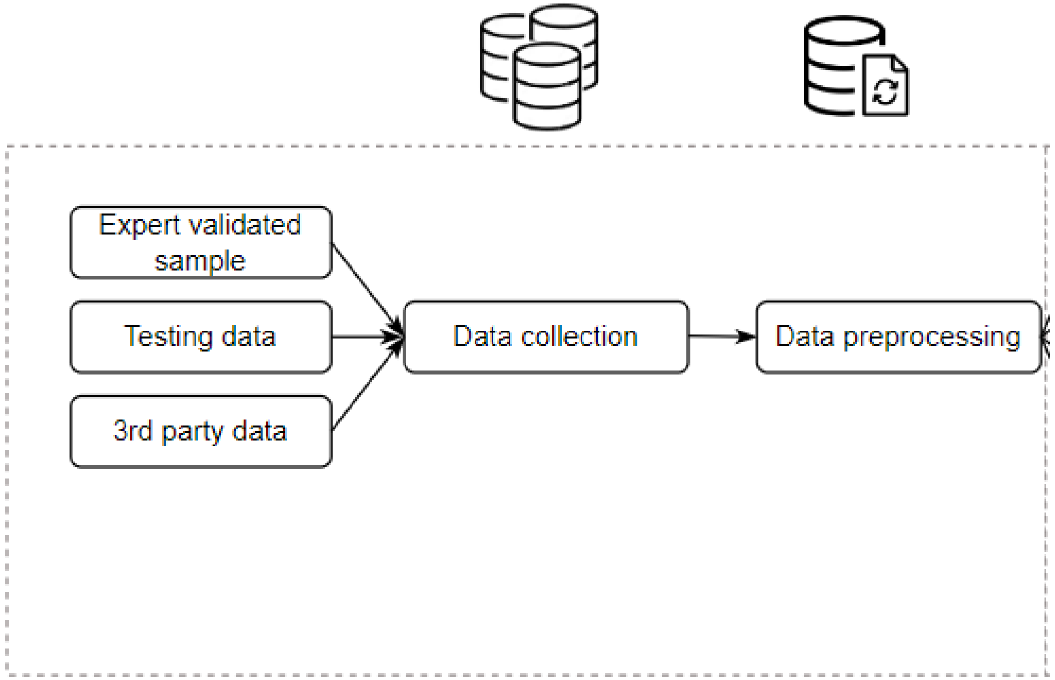
$$VIF_i = \frac{1}{1 - R_i^2}$$



| Feature | VIF |
|---|---|
| Darbo aprašas_Bendra įrenginio būklė | 2.22 |
| Darbo aprašas_Dangtelio būklės nustatymas | 2.42 |
| Darbo aprašas_Detalių užsakymas remontui | 1.63 |
| Darbo aprašas_Pirminė įrenginio būklė | 1.81 |
| Chargeable | 1.31 |
| Battery Cycle Count | 2.40 |
| battery_health | 2.42 |
| battery_lifetime | 1.44 |
| Battery Temperature | 1.68 |

- Visual identification - plotting pairwise scatterplots between independent variables can help visualize the relationships between them.

# Case Study: Automated Refurbished Smartphones Remaining Value Estimation Framework

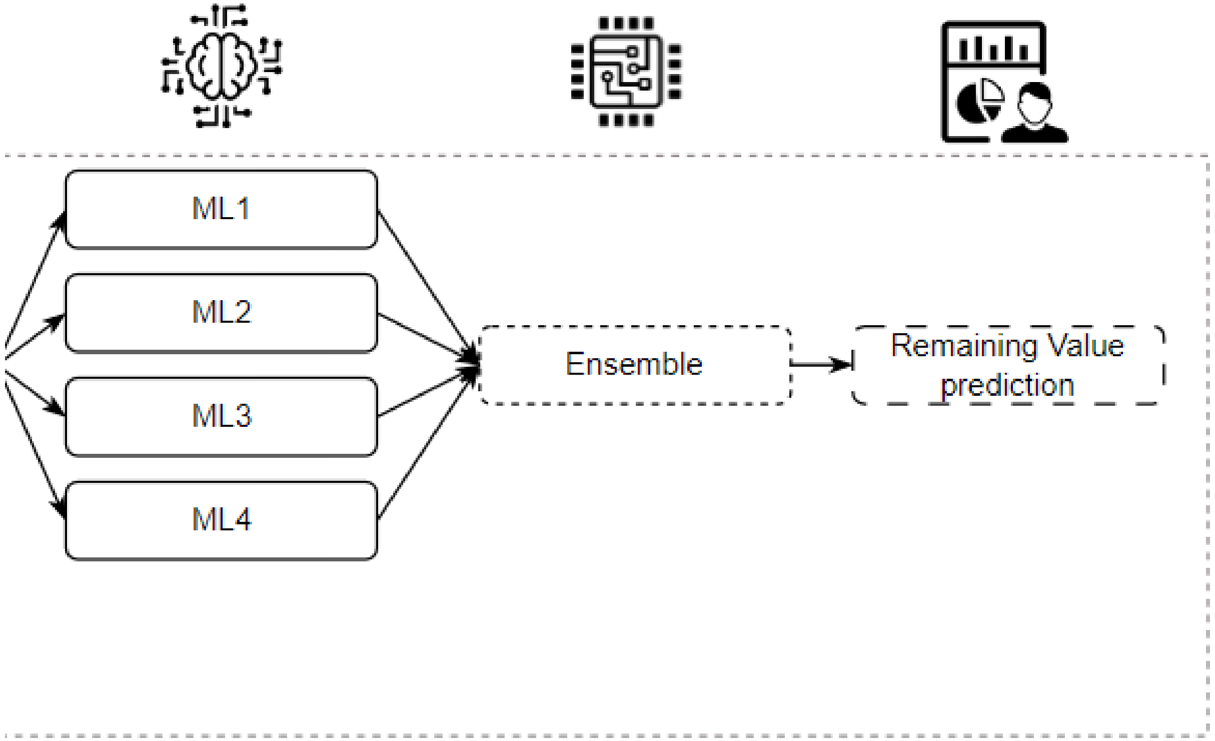


- ➢ Semi-supervised model evaluation after hyperparameters tunning.

| Model | Labelled sample $R^2$ |
|---|---|
| Linear Regression | 0,56 |
| Random Forest | 0,8 |
| XgBoost | 0,9 |
| AdaBoost | 0,82 |

- ➢ XgBoost model selected for the pseudo labels generation.

# Case Study: Data preprocessing



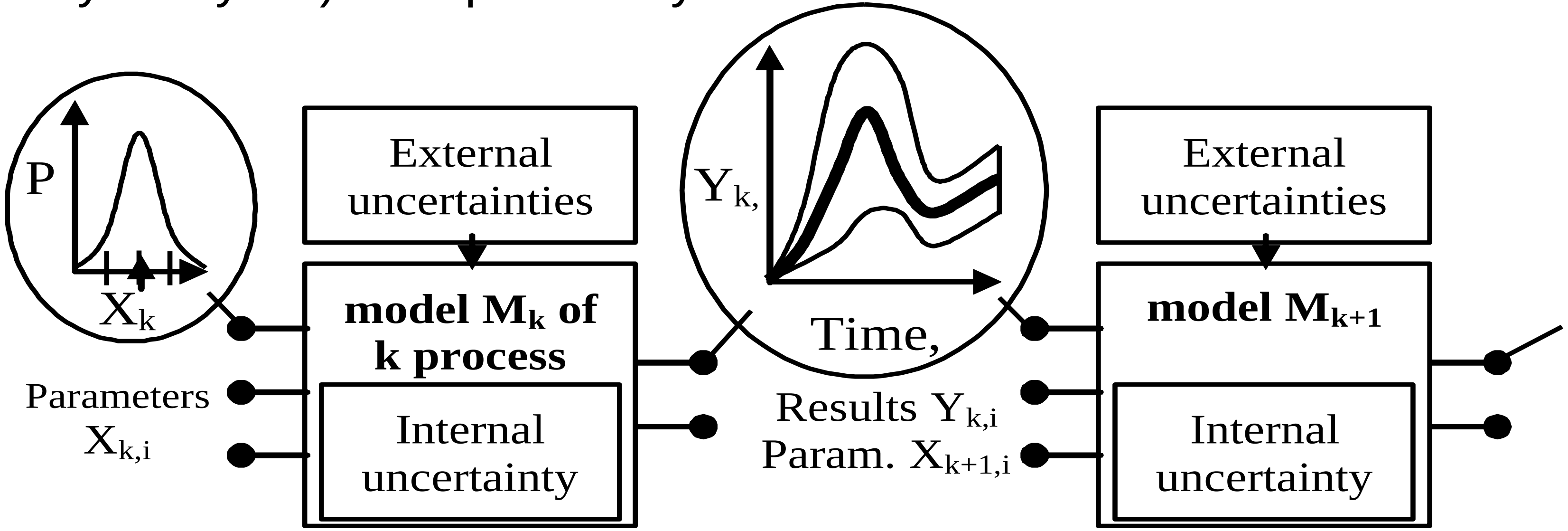
- Pseudo labeled values used as input in Supervised learning algorithms.

- Single model and meta model results:

| Model | $R^2$ |
|---|---|
| Linear Regression | 0.72 |
| Random Forest | 0.84 |
| XgBoost | 0.87 |
| **Meta Model** | 0.89 |

# Uncertainty Estimation in the Modelling Process

- The approach suggested for uncertainty and sensitivity analysis is based on well-established concepts and tools (e. g. SimLAB, SUSA - Software System for Uncertainty and Sensitivity Analyses) from probability calculus and statistics.

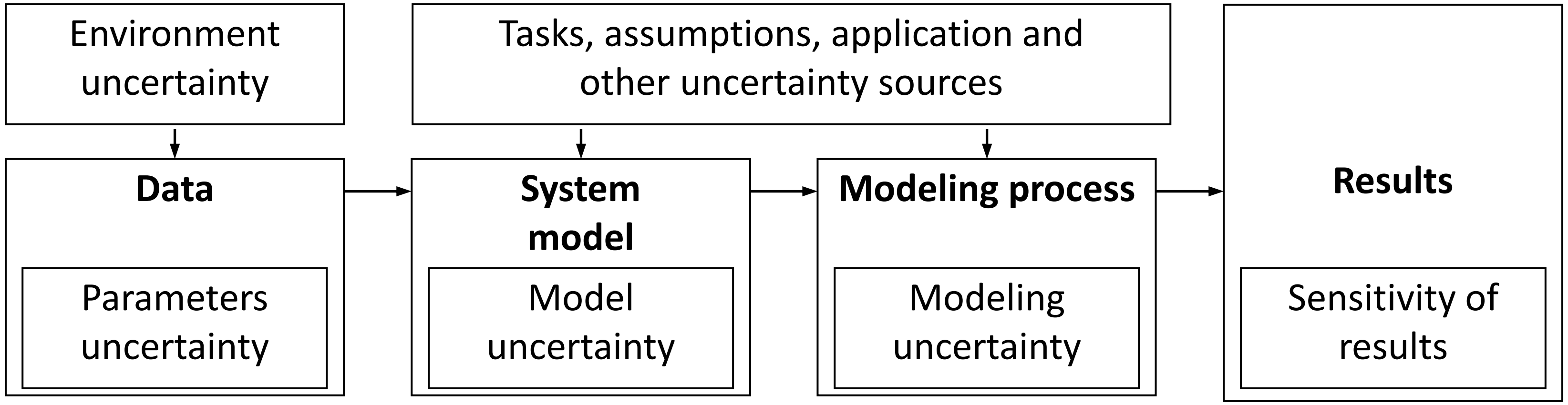


Distribution estimation for process parameters

- It requires identification of the potentially important contributors to the uncertainty of the results and the quantification of the respective state of knowledge by subjective probability distributions.
- Such a distribution expresses how well an uncertain parameter of the model application is known.

# Uncertainty/Sensitivity Estimation/Analysis

- The aim of sensitivity analysis is to identify the main contributors to the possible variability of results.
- Sensitivity analysis is performed in connection with uncertainty analysis in order to see the combined influence of all the potentially important uncertainties on the result.



**Uncertainty and sensitivity estimation process**

- In order to rank uncertainties according to their contribution to output uncertainty, standardized regression coefficients (SRCs) might be chosen from the many other measures available.
- They are capable to indicate the direction of the contribution. Additionally, using sample-based method the different correlation ratios are computed/compared. +Variance-based methods (FAST, Sobol, etc.).

# Conclusion

- Increased consumer awareness along with advances in Machine Learning techniques, enable decision-makers with data-driven insights for informed decision-making.

- With the absence of a gold standard or labeled data for model validation, data preprocessing becomes a critical process.

- Semi-supervised learning methods enable the application of supervised learning approaches for accurate remaining value estimation, effectively reducing data labeling costs.

- Combination of multiple regression models into a single one improves the model performance of Remaining Value Estimation $R^2$ from 0.87 to 0.88.

HiTEc meeting &
Workshop on Complex data in
Econometrics and Statistics

# Questions?

**Ieva Dundulienė,** ieva.dirdaite@ktu.edu

**Robertas Alzbutas,** robertas.alzbutas@ktu.lt