Date: November 9, 2018

Technical Report

De : Paul Kowalczyk

À: Jean-Yves Delannoy
Alessio Tamburro

Copie :
Ref :

Pages : 21

Machine Learning
Final Report.

Informatics-based statistical learning approaches can be a promising alternative to quantum mechanical computations for calculating bandgaps. With this report we demonstrate a systematic feature-engineering approach and a robust learning framework for efficient and accurate predictions of electronic bandgaps of sulfides. The models developed are validated and tested using data science best practices. Following a feature-engineering protocol that included (1) removing features with zero variance, (2) removing highly correlated features, and (3) removing features that result from linear combinations of other features, six machine learning algorithms were tested. These algorithms were: multiple linear regression, partial least squares, support vector machines, k-nearest neighbors, boosted trees, and random forests. Models were built using each algorithm. A training set of 1593 sulfide band gap energies used for model construction was retrieved from the Materials Project. The predictive accuracy of each model was evaluated using a test set of 396 sulfide band gap energies, also retrieved from the Materials Project. Using the adjusted R^2 of the test set results, the boosted trees machine learning algorithm was selected as the best model (adjusted $R^2 = 0.8953$).

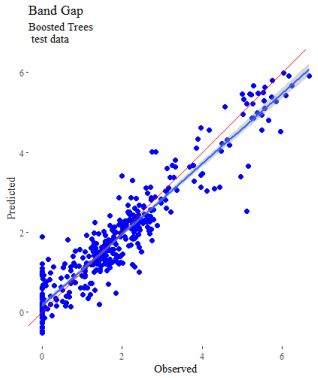

Figure 1: Boosted Trees: predicted vs observed. Adjusted $R^2 = 0.8953$

Contents

Chapter 1

Organization of this final report

Four chapters follow:

Chapter 2. Introduction This chapter introduces machine learning, describing both the construction of machine learning models and machine learning workflows.

Chapter 3. Materials and Methods This chapter describes

1. the data used in this study,
2. the features (descriptors) calculated for each compound, and
3. the machine learning algorithms employed.

Chapter 4. Results and Discussion Each model built was evaluated with a set of test data. The results of each model's predictive performance on this test data is presented in this chapter. The selection of an optimal model is discussed.

Chapter 5. Conclusion The report concludes with a discussion of how this work may be generally applied.

Chapter 2

Introduction

Machine learning models for materials properties are constructed from three parts

1. a training set,
2. a set of attributes describing each material, and
3. machine learning algorithms mapping the attributes to properties.

A training set

Massive open-access databases of computed/predicted materials properties (including electronic structure, thermodynamic, and structural properties) are now available. [AFLOWLIB.ORG, Materials Project, Computational Materials Repository] Methods are sought to efficiently extract knowledge and mine trends out of these materials big-data repositories.

Given prior knowledge - in terms of high quality data on a given property of interest for a limited set of material candidates within a well defined chemical space - informatics based statistical learning approaches lead to efficient pathways to make high-fidelity predictions on new compounds within the target chemical space.

Our goal is the construction of validated statistical learning models for the prediction of bandgaps of sulfides. The models would establish a mathematical relationship (a *mapping*) between the bandgap of material i residing in the predefined chemical space, and an Ω-dimensional (Ω-D) feature vector f_i of the material i.

A set of attributes describing each material

An attribute (also referred to as a feature; used interchangeably in this report) is an individual measurable property or characteristic of a object being studied. Choosing informative, discriminating and independent attributes is a crucial step for effective algorithms in pattern recognition, classification and regression.

The concept of "feature" is related to that of explanatory variable used in statistical techniques such as linear regression.

Features may be *observed*, e.g., date of experiment, location of experiment, environmental conditions. Features may also be *calculated*, e.g., molecular weight, number of atoms in a molecule, presence/absence of particular substructures.

The initial set of raw features can be redundant and too large to be managed. Therefore, a preliminary step in many applications of machine learning and pattern recognition consists of selecting a subset of features, or constructing a new and reduced set of features to facilitate learning, and to improve generalization and interpretability[citation needed].

Extracting or selecting features is a combination of art and science; developing systems to do so is known as feature engineering. It requires the experimentation of multiple possibilities and the combination of automated techniques with the intuition and knowledge of the domain expert. Automating this process is feature learning, where a machine not only uses features for learning, but learns the features itself.

Feature engineering is the process of using domain knowledge of the data to create features that make machine learning algorithms work.

Machine learning algorithms mapping the attributes to properties

Machine learning algorithms can be separated into two broad classes: *supervised* and *unsupervised* learning. In both of these classes, the algorithm has access to a set of observations known as *training data*, the collection of known input ($\mathbf{X}$) and output (y) values, which may be generated through observations or controlled experiments. The goal of the scientist is to use such training data, as well as any other prior knowledge, to identify a function that is able to predict the output value for a new set of input data accurately. In supervised learning, the training data consists of a set of *input values* (e.g., the structures of different materials) as well as a corresponding set of *output values* (e.g., materials property values). With these training data, the machine learning algorithm tries to identify a function that can make accurate predictions about the output values that will be associated with new input values. If the output values y form a continuous range (e.g., melting points), then the process of searching for a function is known as *regression*. If the allowed output values form a discrete set (e.g., space groups), the process is then known as *classification*. In unsupervised learning, there are no output values in the training data, and the goal is to identify patterns in the input values.

Our approach to machine learning is based on the **cr**oss-**i**ndustry **s**tandard **p**rocess for **d**ata **m**ining (CRISP-DM). A process diagram of this process is presented in Figure 2.1.

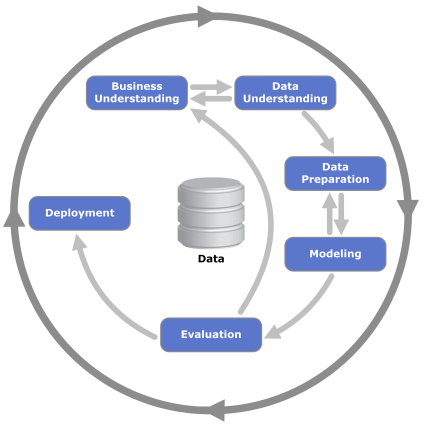

Figure 2.1: Process diagram of the CRISP-DM process.

CRISP-DM breaks the process of data mining into six major phases. The sequence of the phases is not strict, and moving back and forth between different phases is always required. The arrows in the process diagram indicate the most important and frequent dependencies between phases. The outer circle in the diagram symbolizes the cyclic nature of data mining itself. A data mining process continues after a solution has been deployed. The lessons learned during the process can trigger new, often more focused business questions, and subsequent data mining processes will benefit from the experiences of previous ones.

During the **business understanding** phase business objectives and requirements are investigated; decisions are made as to whether data mining can be applied to meet them; and determinations are made regarding the data that can be collected to build a deployable model. During the **data understanding** phase an initial dataset is established and studied to see whether it is suitable for further processing. Insights gained during this stage may lead to collecting new data or, in some cases, reconsidering the utility of a data mining campaign.

An actual example of **data preparation**, **modeling**, and **evaluation** is presented throughout this report. Preparation involves preprocessing the raw data so that machine learning algorithms can produce a model. During the modeling phase numerous algorithms are used to construct relationships between the features ($\mathbf{X}$) and outcomes (y). The importance of the **evaluation** phase cannot be overstated. Only those

models showing high accuracy and predictive ability merit further consideration. It can happen that no acceptable model is identified.

One enters the **deployment** phase once an acceptable model has been identified. Here, the model is democratized, i.e., made generally available. Discussions of model deployment are outside the scope of this report.

Chapter 3

Materials and Methods

3.1 Data Set

Data for this report were retrieved from the Materials Project. Nine properties were retrieved for each entry:

pretty_formula a nice formula where the element amounts are normalized

unit_cell_formula the full explicit formula for the unit cell

energy calculated VASP energy for structure (the Vienna Ab-initio Simulation Package (VASP) is a software package that uses density functional theory to solve the quantum problem for materials)

band_gap the calculated band gap

e_above_hull calculated energy above convex hull for structure

nelements the number of elements in the material

spacegroup.crystal_system the crystal system for the space group

spacegroup.number the number for the space group

spacegroup.symbol the symbol for the space group

A total of 83989 entries were retrieved.

Only energetically stable compounds were considered in this workflow. Consequently, only those compounds with e_above_hull $\leq 10^{-9}$ were retained. 25370 entries had an e_above_hull $\leq 10^{-9}$.

Further, only sulfides were considered in this workflow. There were 1989 sulfides amongst the 25370 stable compounds.

A sampling of the retrieved data is presented in Table 3.1.

composition	band gap	unit cell formula
BaCdGeS4	2.7009	'S': 32.0, 'Ge': 8.0, 'Cd': 8.0, 'Ba': 8.0
AgBiSCl2	1.0687	'Ag': 2.0, 'Bi': 2.0, 'S': 2.0, 'Cl': 4.0
NbCu3S4	1.6574	'Nb': 1.0, 'Cu': 3.0, 'S': 4.0
K2NdP2S7	2.3533	'P': 8.0, 'S': 28.0, 'K': 8.0, 'Nd': 4.0
Cu3SbS4	0.0000	'S': 4.0, 'Cu': 3.0, 'Sb': 1.0
BaCu2SnS4	0.3869	'S': 12.0, 'Cu': 6.0, 'Sn': 3.0, 'Ba': 3.0
KSb(PS3)2	2.2673	'K': 2.0, 'Sb': 2.0, 'P': 4.0, 'S': 12.0
ThAsS	0.0000	'Th': 2.0, 'As': 2.0, 'S': 2.0
K3Cu2(BiS2)5	0.5758	'K': 6.0, 'Cu': 4.0, 'Bi': 10.0, 'S': 20.0
MgSO4	5.4889	'O': 8.0, 'Mg': 2.0, 'S': 2.0

Table 3.1: Sampling of data retrieved from the Materials Project.

3.2 Features

Features were calculated using Magpie [**M**aterials-**Ag**nostic **P**latform for **I**nformatics and **E**xploration] (REFERENCE A general - purpose machine learning framework for predicting properties of inorganic materials. Logan Ward, Ankit Agrawal, Alok Choudhary & Christopher Wolverton. npj Computational Materials volume 2, Article number 16028 (2016)).

Magpie allows for the calculation of an expansive set of attributes that can be used for materials with any number of constituent elements. This set is broad enough to capture a sufficiently diverse range of physical/chemical properties to be used to create accurate models for many materials problems. In total, a set of 145 attributes are calculated. These attributes fall into four distinct categories:

1. **Stoichiometric attributes** that depend only on the fractions of elements present and not what those elements actually are. These include the number of elements present in the compound and several Lp norms of the fractions.

2. **Elemental property statistics**, which are defined as the mean, mean absolute deviation, range, minimum, maximum and mode of 22 different elemental properties. This category includes attributes such as the maximum row on periodic table, average atomic number and the range of atomic radii between all elements present in the material.

3. **Electronic structure attributes**, which are the average fraction of electrons from the s, p, d and f valence shells between all present elements.

4. **Ionic compound attributes** that include whether it is possible to form an ionic compound assuming all elements are present in a single oxidation state, and two adaptations of the fractional 'ionic character' of a compound based on an electronegativity-based measure.

3.3 Classifiers

Chapter 4

Results and Discussion

4.1 Building the training set and the test set

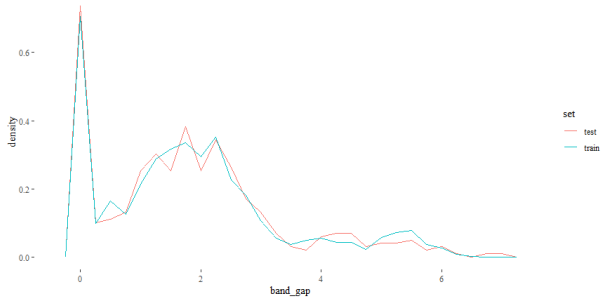

Figure 4.1: Density distributions for bandgap energies in the training and test sets.

4.2 Feature Engineering

4.2.1 Zero variance features

Seven of the calculated features for the training set have zero variance (i.e., are single-valued). These features are max_NsValence, min_NdValence, min_NfValence, min_NsUnfilled, min_NdUnfilled, and min_NfUnfilled, min_GSmagmom. These features are removed from both the training set and the test set.

4.2.2 Feature Correlation

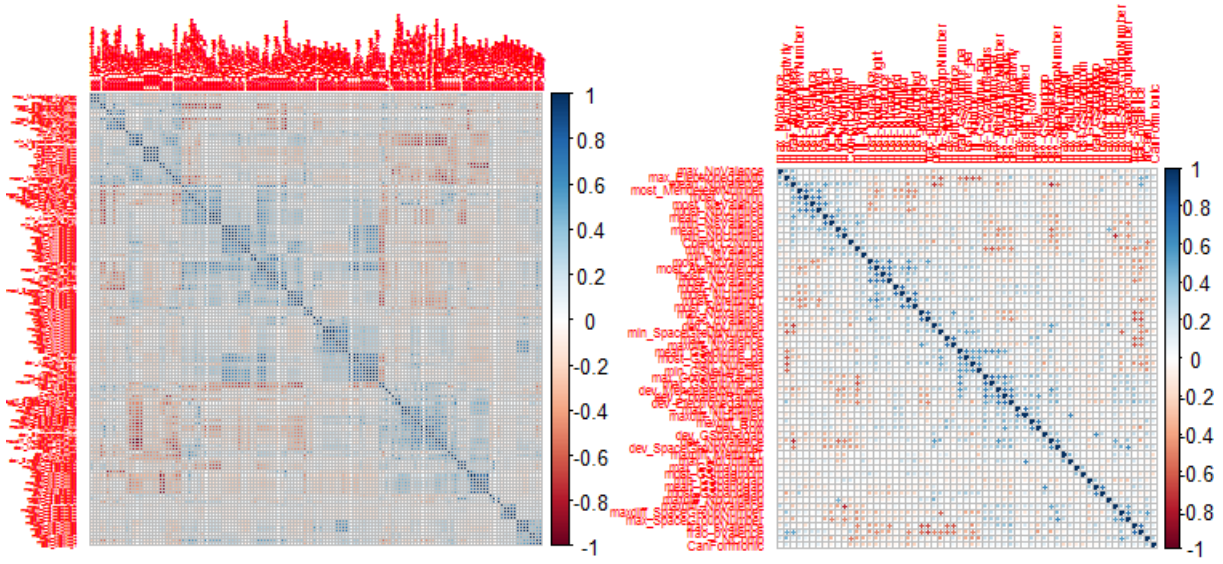

Figure 4.2: Graphical display of the feature correlation matrix before (left) and after (right) removing highly correlated features.

4.3 Machine Learning Models

4.3.1 Multiple Linear Regression

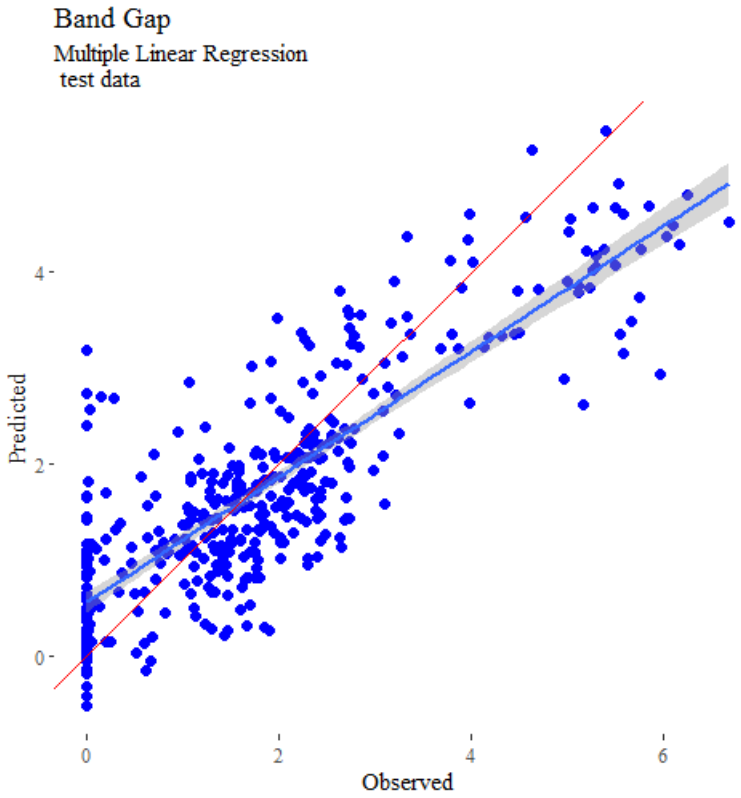

Figure 4.3: Multiple Linear Regression: predicted vs observed. Adjusted $R^2 = 0.6947$

4.3.2 Partial Least Squares

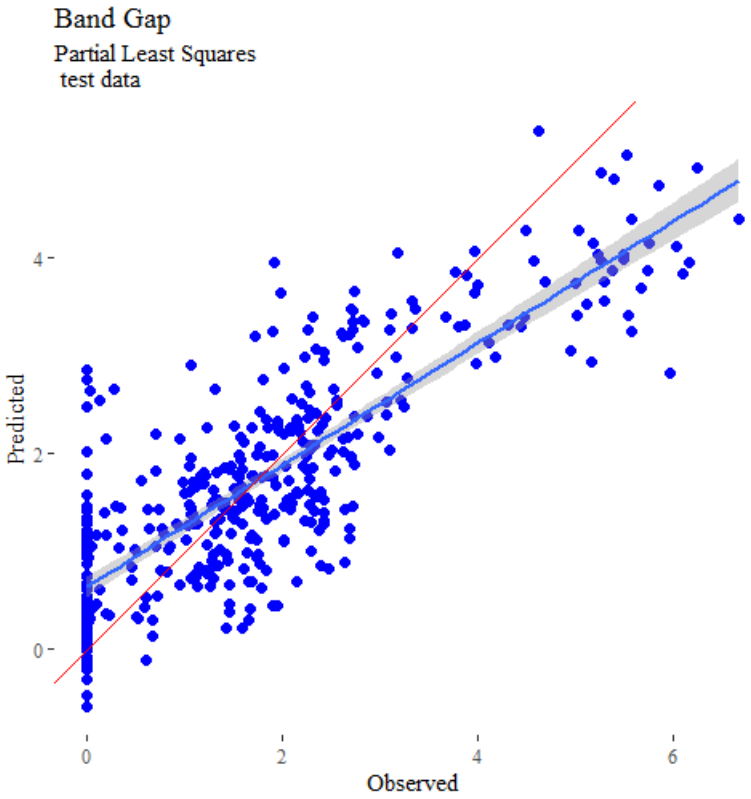

Figure 4.4: Partial Least Squares: predicted vs observed. Adjusted $R^2 = 0.6652$

4.3.3 Support Vector Machine

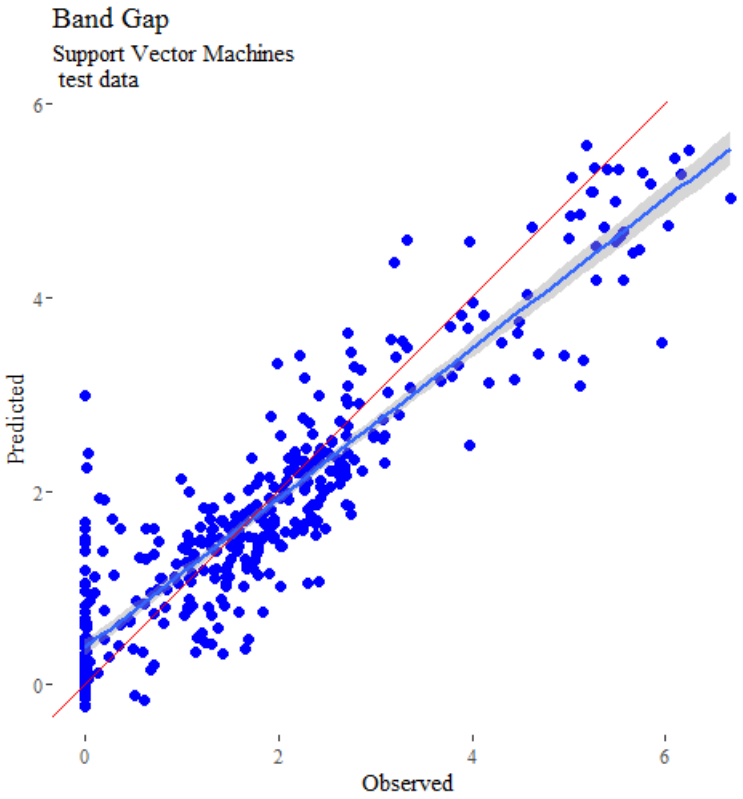

Figure 4.5: Support Vector Machine: predicted vs observed. Adjusted $R^2 = 0.8294$

4.3.4 K-Nearest Neighbors

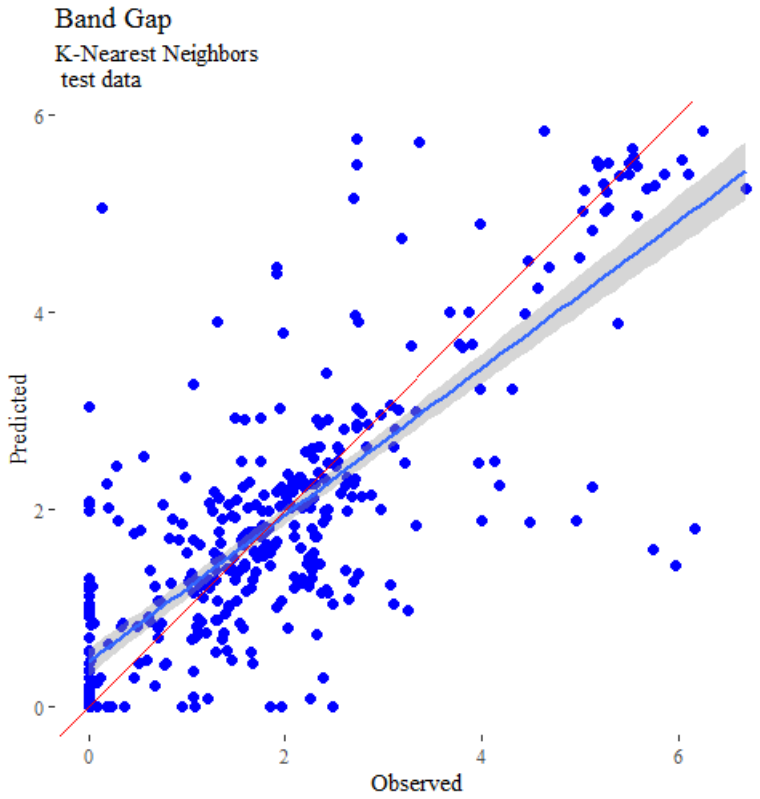

Figure 4.6: K-Nearest Neighbors: predicted vs observed. Adjusted $R^2 = 0.6207$

4.3.5 Random Forest

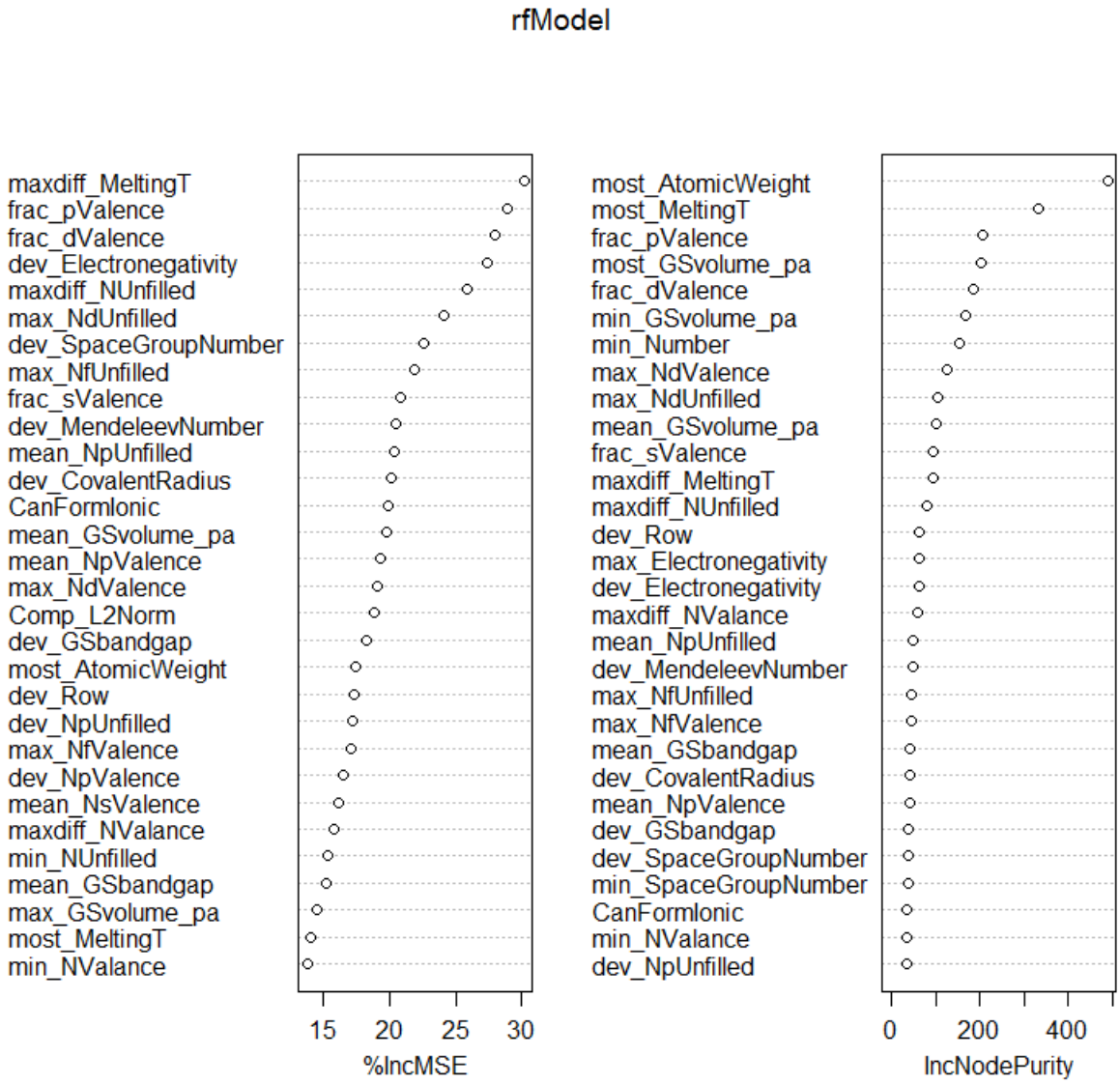

Figure 4.7: Random Forest: variable importance plots.

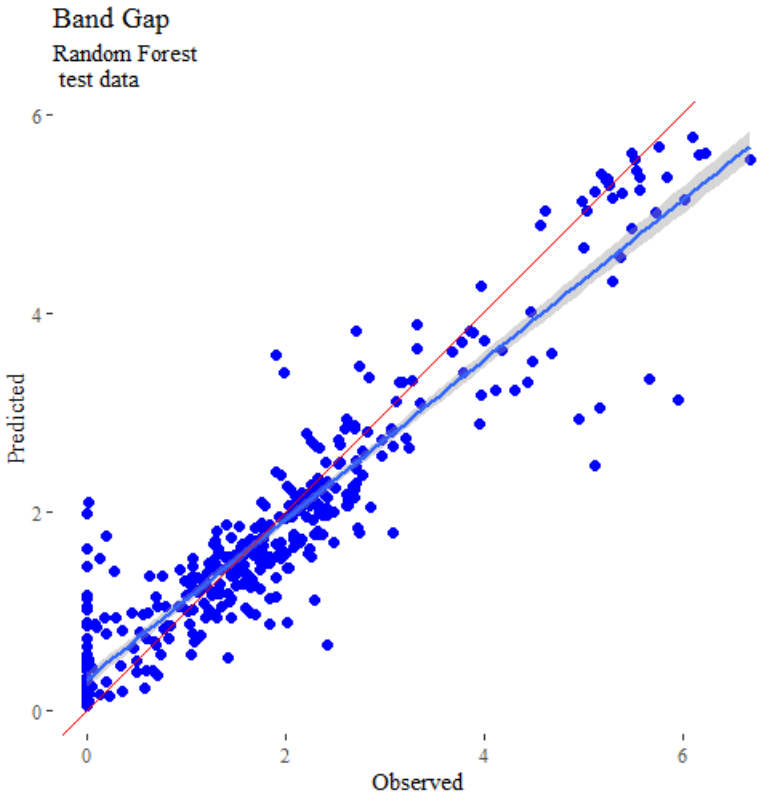

Figure 4.8: Random Forest: predicted vs observed. Adjusted $R^2 = 0.8746$

.

4.3.6 Boosted Trees

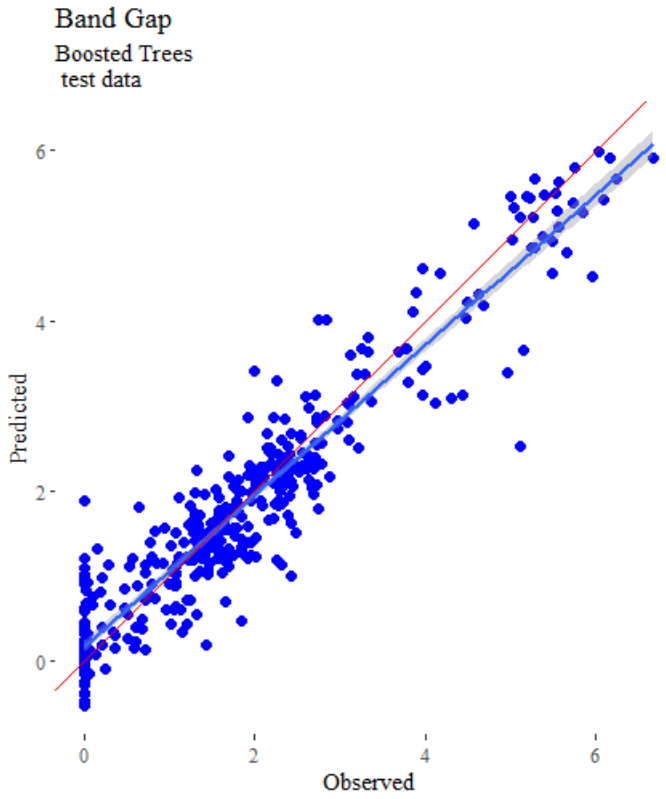

Figure 4.9: Boosted Trees: predicted vs observed. Adjusted $R^2 = 0.8953$

.

Chapter 5

Conclusion

Machine Learning Method	**Adjusted R^2**
Multiple Linear Regression	0.6947
Partial Least Squares	0.6652
Support Vector Machine	0.8294
K-Nearest Neighbors	0.6207
Boosted Trees	0.8953
Random Forest	0.8746

Table 5.1: Summary of machine learning experiments.

FICHE RESUME/BIBLIOGRAPHICAL FORM DOCUMENTUM

SERVICE	Laboratory Group	SM@RT
Type de document	Document type	Rapport de Projet
Date	Application Date	Decembre 2014
TITRE en Anglais	English Title	COMPNANOCOMP FP7 Project : Final Report
TITRE en Français	French Title	Projet FP7 COMPNANOCOMP : Rapport Final
Entreprise	Enterprises/Sponsors	R&I/AIO/Advanced Materials Platform
Auteur(s)	Authors	Jean Yves DELANNOY
PROJET	Projects	COMPNANOCOMP
Collaborateurs	Collaborators	Cédric Feral-Martin Aurélie Papon Magali Fontana Olivier Sanseau
N^0 d'affaire	Business code	
RESUME Anglais		The objective of this document is to offer a "digest" of the main results obtained in this project that aims at the development of multiscale simulation methodology and software for predicting the morphology (spatial distribution and state of aggregation of nanoparticles), thermal (glass temperature), mechanical (viscoelastic storage and loss moduli, plasticity, fracture toughness and compression strength), electrical and optical properties of soft and hard polymer matrix nanocomposites from the atomic-level characteristics of their constituent nanoparticles and macromolecules and from the processing conditions used in their preparation. The document gives a short summary of the results obtained within this project by Solvay and its partners. It also emphasizes the interest for Solvay R&I of the developments obtained.

RESUME Français		L'objectif de ce rapport est de fournir un résumé des points essentiels développés dans le rapport final du projet européen FP7 COMPNANOCOMP dont le but est de développer une méthodologie de simulation multi-échelle permettant de prédire la morphologie (distribution spatiale et état d'agrégation des nanoparticules), et les propriétés thermiques, mécaniques optiques et électriques de nanocomposites de polymères. Ce travail s'effectue sur la base des caractéristiques atomiques des constituants du matériaux et doit prendre en compte les conditions du procédé ayant permis sa réalisation. Ce document donne une vision des résultats obtenus par Solvay et ses collaborateurs et met en avant l'intérêt pour le groupe des développements effectués.
Mots Clés Anglais	English Keywords	RUBBER, REINFORCEMENT, SILICA, COUPLING, MODELING, MORPHOLOGY, FP7
Mots Clés Français		CAOUTCHOUC ; RENFORT ; SILICE ; COUPLAGE ; MODELISATION ; MORPHOLOGIE, FP7
RNCAS	RNCAS	
Destinataires	Addressees	
CONFIDENTIEL	Confidential	YES