IDENTIFY FRAUD FROM ENRON EMAIL

Summarize for us the goal of this project and how machine learning is useful in trying to accomplish it. As part of your answer, give some background on the dataset and how it can be used to answer the project question. Were there any outliers in the data when you got it, and how did you handle those? [relevant rubric items: "data exploration", "outlier investigation"]

What features did you end up using in your POI identifier, and what selection process did you use to pick them? Did you have to do any scaling? Why or why not? As part of the assignment, you should attempt to engineer your own feature that does not come ready-made in the dataset – explain what feature you tried to make, and the rationale behind it. (You do not necessarily have to use it in the final analysis, only engineer and test it.) In your feature selection step, if you used an algorithm like a decision tree, please also give the feature importances of the features that you use, and if you used an automated feature selection function like SelectKBest, please report the feature scores and reasons for your choice of parameter values. [relevant rubric items: "create new features", "properly scale features", "intelligently select feature"]

What algorithm did you end up using? What other one(s) did you try? How did model performance differ between algorithms? [relevant rubric item: "pick an algorithm"]

What does it mean to tune the parameters of an algorithm, and what can happen if you don,Äôt do this well? How did you tune the parameters of your particular algorithm? (Some algorithms do not have parameters that you need to tune – if this is the case for the one you picked, identify and briefly explain how you would have done it for the model that was not your final choice or a different model that does utilize parameter tuning, e.g. a decision tree classifier). [relevant rubric item: "tune the algorithm"]

What is validation, and what's a classic mistake you can make if you do it wrong? How did you validate your analysis? [relevant rubric item: "validation strategy"]

Give at least 2 evaluation metrics and your average performance for each of them. Explain an interpretation of your metrics that says something human-understandable about your algorithm's performance. [relevant rubric item: "usage of evaluation metrics"]

Techniques and algorithms that have stemmed from the field of *machine learning* have indeed now become a powerful tool for the analysis of complex and large data, successfully assisting scientists in numerous breakthroughs of various fields of science and technology.

Public and famous examples include the use of boosted decision trees in the statistical analysis that led to the detection of the Higgs boson at CERN [**?**], the use of random forests for human pose detection in the Microsoft Kinect [**?**] or the implementation of various machine learning techniques for building the IBM Watson system [**?**], capable to compete at the human champion level on the American TV quiz show Jeopardy.

(a) First, the theoretical properties and statistical mechanisms that drive the algorithm are still not clearly and entirely understood. Random forests indeed evolved from empirical successes rather than from a sound theory. As such, various parts of the algorithm remain heuristic rather than theoretically motivated. For example, preliminary results have proven the consistency of simplified to very close variants of random forests, but consistency of the original algorithm remains unproven in a general setting.

(b) Second, while the construction process of a single decision tree can easily be described within half a page, implementing this algorithm properly and efficiently remains a challenging task involving issues that are easily overlooked. Unfortunately, implementation details are often omitted in the scientific literature and can often only be found by diving into (unequally documented) existing software implementations. As far as we know, there is indeed no comprehensive survey covering the implementation details of random forests, nor with their respective effects in terms of runtime and space complexity or learning ability.

(c) Third, interpreting the resulting model remains a difficult task, for which even machine learning experts still fail at finely analyzing and uncovering the precise predictive structure learned by the procedure. In particular, despite their extensive use in a wide range of applications, little is still known regarding variable importance measures computed by random forests. Empirical evidence suggests that they are appropriate for identifying relevant variables, but their statistical mechanisms and properties are still far from being understood.

0.1 OUTLINE AND CONTRIBUTIONS

Part **??** of this manuscript is first dedicated to a thorough treatment of decision trees and forests of randomized trees. We begin in Chapter **??** by outlining fundamental concepts of machine learning, and then proceed in Chapters **??** and **??** with a comprehensive review of the algorithms at the core of decision trees and random forests. We discuss the learning capabilities of these models and carefully study all parts of the algorithm and their complementary effects. In particular, Chapter **??** includes original contributions on the bias-variance analysis of ensemble methods, highlighting how randomization can help improve performance. Chapter **??** concludes this first part with an original space and time complexity analysis of random forests (and their variants), along with an in-depth discussion of implementation details, as contributed within the open source Scikit-Learn library. Overall, Part **??** therefore presents a comprehensive review of previous work on random forests, including some original contributions both from a theoretical and practical point of view.