# Book Proposal

**Proposed Book Title:** Fraud Detection with Machine Learning

**Proposed Book Subtitle:**

**Author:** Chris Kuo, Ph.D.

**Author titles and affiliations:**

* Lecturer, Master of Science Applied Analytics @ Columbia University
* Director, Data Science, The Hartford Insurance Company

**Preferred mailing address:**

* 69-13 67 Pl., Glendale, NY 11385

**Preferred phone number:**

* 617-901-3882

**Preferred Email address:**

* ck2869@columbia.edu

**Author Platform details**

*Author biography:*

[*http://sps.columbia.edu/applied-analytics/master-of-science-in-applied-analytics/faculty/chris-kuo*](http://sps.columbia.edu/applied-analytics/master-of-science-in-applied-analytics/faculty/chris-kuo)

*LinkedIn profile:*

[*https://www.linkedin.com/in/chriskuo/*](https://www.linkedin.com/in/chriskuo/)

*Author Web Twitter:*

*N/A*

*Do you have a website/blog? (if so, provide URL)*

*N/A*

[*https://github.com/ChrisKuoColumbiaU/FraudDetection*](https://github.com/ChrisKuoColumbiaU/FraudDetection)

*Do you participate in meetups, conferences or similar events? If so, which ones?*

* *AWS Loft in NYC*
* *NYC Meetups*
* *Joint statistical meetings*
* *Conferences @ Columbia U. School of Professional Studies*

*Do you give presentations? Are any available online?*

* *At working places*
* *No online content*

*Why are you the best person to write this book?*

* *17+ progressive leadership experiences in data science in insurance, risk segmentation, claims, fraud detection, litigation management, customer analytics, retail management.*
* *15+ years teaching experiences with average rating 4.5/5.0*
* *Published in peer-reviewed journal in economics & management*

## Book Summary:

*In one sentence, tell us why the audience will want to buy your book.*

*A book marrying the fraud practices with code examples for data scientists & fraud detection practitioners*

*Summarize what the book is about, like you would pitch it to a potential reader on the back cover.*

There are many books on machine learning and artificial intelligence for Ph.D. students in computer science. While turning practical problems into elegant mathematical solutions, there is a need for the new data scientists to survey the wide range of fraudulent behaviors and consequences. On the other hand, fraud detection practitioners need a suite of new machine learning techniques to fight against fraud effectively. In this book we attempt to cover the best practices in fraud detection, as well as the cutting-edge machine learning techniques. The author draws examples from insurance, financial services and retail industries since he has served in these industries for 17+ years. We hope this book will help data scientists, fraud detection analysts and practitioners at special investigation units (SIU) in fighting fraud and achieve efficiency.

*What makes your book unique in the marketplace?*

This book is about the uses of machine learning techniques in fraud detection. There are books in fraud detection, but few that feature all the different aspects covered in this book. This book discusses the trajectory from machine learning to insight to decision. Each chapter is introduced by a nontechnical explanation of the machine learning concept, followed by code examples.

<Journal research>

This book presents what research has been done in fraud detection. Research projects range from healthcare, credit card, insurance, payments and operation. This book selects cited researches. Readers will be inspired to apply solutions to similar type of problems. This book also includes datasets from Kaggle competitions and public domains for readers to test.

<Code examples>

This book shows the readers the basic building blocks in Python and R for the selected machine learning techniques. By learning the code examples, you will be hands-on with the step-by-step actions the code takes to complete the tasks.

<Use cases>

At the end of each major session this book challenges the data scientists with the write-up of a use case for fraud detection.

## Technology Summary:

*How would you characterize the technology’s stage of development? (Put an X in the column next to the stage that best applies.)*

| Answer | Stage | Description |
| --- | --- | --- |
|  | Emerging | This is a new and cutting-edge technology that is still in the experimental/lab stage. |
| *X* | Developing | A technology that is gaining momentum and has a lot of activity but is not a clear standard. High growth, but multiple competitors on the horizon. |
|  | Mature | An established technology with leading market/mind share. Strong or moderate growth. |
|  | Declining | An old titan that may still be important and useful, but there is no further expected development work or growth. Alternative approaches threaten to make it obsolete. |

*Briefly explain the technology and why it is important.*

*As fraudsters advance in technology and scale, we need more machine learning techniques to detect earlier and more accurately. Machine learning offers a combination of predictive and behavioral analytics as well as social network analysis to uncover suspicious patterns to improve detection accuracy. The benefits of machine learning include:*

* *Making decisions in-time: Sophisticated analytics should also be built to provide speed and eases business decisions.*
* *Finding new patterns: Machine learning techniques can uncover new patterns to detect changes in fraud behaviors.*
* *Less disruption to regular claims: Genuine claims need to be honored and fraudulent claims need to be deterred. Machine learning aims at lower false positive rate to yield less interruption to regular claims.*
* *Reducing operation cost: a reduction of the payout on fraudulent claims will increase the operation efficiency.*

## Audience:

*Explain who the primary audience is for your book.*

* This book will be used for the course “Fraud detection with Python and R” for the Applied Analytics at Columbia University in 2018 summer.
* This book is for new data scientists to survey the wide range of fraudulent behaviors and consequences.
* This book equips the fraud detection practitioners with a suite of new machine learning techniques to fight against fraud effectively.

*What professional positions does this audience hold?*

* Lecturer, Master of Science Applied Analytics @ Columbia University
* Director, Data Science, The Hartford Insurance Company

*What knowledge do you assume of this audience?*

* Basic Python and R programming

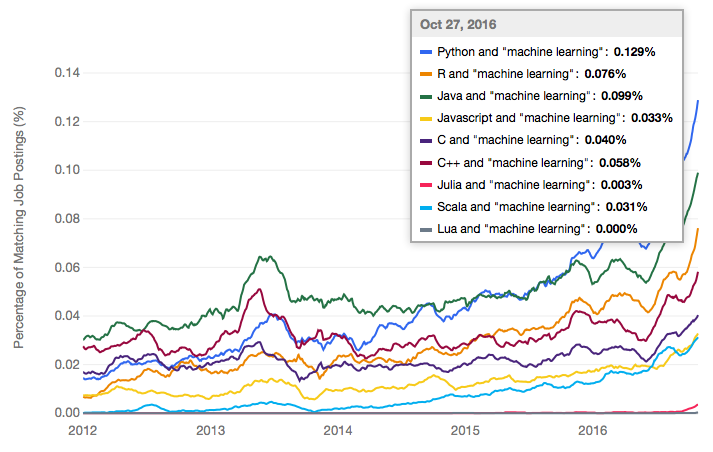
*What books can you assume they have read?*

* *Introductory books to Machine Learning*

*What skills can you assume they have mastered?*

* *Python and R installation*
* *Microsoft office*

*Please estimate as best you can how many people will use this technology? Please state any applicable statistics (e.g., web searches, web site traffic, blogs) indicating market use or market growth.*

**

*https://www.kdnuggets.com/2017/01/most-popular-language-machine-learning-data-science.html*

*Use the following table to describe how the audience for your book typically gets information and where it looks for guidance and leadership (list top five choices).*

|  |  |
| --- | --- |
| What web sites or blogs do they read? | * Code examples on Github * Scikit-learn: <http://scikit-learn.org/stable/index.html> * https://cran.r-project.org/ |
| What publications do they read (e.g., magazines, journals, newspapers)? | * Insurance journal: <https://www.insurancejournal.com/> * Property Casualty 360: <http://www.propertycasualty360.com/> * New England Journal of Medicine: http://www.nejm.org/doi/full/10.1056/NEJMsb1706645#t=article * WSJ.com |
| What conferences do they attend? | * AWS summit * O’Reilly Artificial Intelligence Conference |
| To what professional organizations do they belong? | * AWS summit * Association of Certified Fraud Examiners (ACFE) * Chartered Property Casualty Underwriting (CPCU) |
| Who are the leaders and key influencers in the field who would review or endorse your book? | * Special Investigation Unit (SIU) managers * Head of data science @ companies * Academic professors * Research scientists in healthcare * Chief data scientists/Chief data officers in healthcare |

## Key Topic Coverage:

*What are the top five topics that will be covered in the book? Why are they the top five?*

*A good feature engineering enables a good fraud detection model. Feature engineering also depends on the knowledge for the subject. So this book presents feature engineering for fraud detection in credit card, healthcare and textual data.*

*This book covers supervised learning and unsupervised learning. For the supervised learning this book provide solutions to extreme imbalanced data, which is common in fraud detection.*

*This book covers graph-based techniques to detect fraud ring.*

*The 5 topics are listed below:*

1. *Feature engineering for fraud detection*
   * *Features for credit card transaction fraud*
   * *Features for healthcare fraud, waste and abuse*
   * *Extracting features from texts*
2. *Sampling techniques for extremely imbalanced data*
   * *What is imbalanced data?*
     + *Why an ROC curve cannot measure well*
   * *Undersampling techniques*
     + *Random undersampling for the majority class*
     + *NearMiss*
     + *Condensed Nearest Neighbor Rule (CNN)*
     + *TomekLinks*
     + *Edited Nearest Neighbor Rule (ENN)*
     + *NeighbourhoodCleaningRule*
     + *ClusterCentroids*
     + *NeighbourhoodCleaningRule*
   * *Oversampling techniques*
     + *Random oversampling for the minority class*
     + *Synthetic Minority Oversampling Technique (SMOTE)*
     + *ADASYN: Adaptive Synthetic Sampling*
3. *Supervised learning techniques for fraud detection*
   * *GBM*
   * *Xgboost*
   * *What is deep learning?*
     + *Deep learning using Keras*
     + *Keras for credit card fraud detection*
4. *Unsupervised learning techniques for fraud detection*
   * *Z-score*
   * *Median absolute deviation (MAD)*
   * *One-class Support Vector Machine (One-class SVM)*
   * *Hidden markov model (HMM)*
   * *Gaussian Mixture Model (GMM)*
   * *Density-based spatial clustering of applications with noise (DBSCAN)*
   * *MeanShift*
   * *Hierarchical clustering*
   * *Self-organizing maps (SOM)*
5. *Graph-based techniques for fraud detection*
   * *Fraud ring example 1: Auto accident insurance claims*
   * *Fraud ring example 2: First-party fraud in retail banking*
   * *Why the "traditional" anomaly detection techniques cannot catch them?*
   * *Graph database: Neo4j*
   * *Using Neo4j to identify fraud ring*

*What problems does this book solve for its users?*

This book describes the challenges in fraud detection in banking, insurance, healthcare as well as other fields. It identifies the right machine learning approaches. The audience will be able to:

* Identify potential applications of machine learning in fraud detection.
* Represent data in feature engineering to yield a better outcome.
* Assess the model quality with the appropriate error metrics.
* Implement these techniques in Python and R.

*List the four or five topics covered or features included that will provide the greatest benefit to readers or will be the most likely to excite them?*

*Listed above.*

*Is there a companion web site?*

*Yes. The code will be made available at*

* [*https://github.com/ChrisKuoColumbiaU/FraudDetection*](https://github.com/ChrisKuoColumbiaU/FraudDetection)

*If so, what do you plan to include on the site?*

*The code examples, and the answer keys.*

*Would you be willing to participate in video offerings as well as workshops and training seminars?*

*To be determined. The book will be the textbook for the course “Fraud Detection with Machine Learning”. Columbia University requires video course content. It is not clear to me at this moment the ownership of the video content*

*Other formats for the content, including video, conference workshops, articles, etc.*

*To be determined for similar reason.*

## Competition:

*What books or online resources compete with this book? Please list the title and author. In each case, how will your book be different or better in timing, content, coverage, approach, or tone? Competition in the market and what makes this an O'Reilly book*

* *The competition is:*
  + *Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques: A Guide to Data Science for Fraud Detection (Wiley and SAS Business Series)*
    - *The fraud analytics process model*
    - *Big data*
    - *Break point/peer group analysis*
    - *Anomaly detection*
    - *Linear/logistic regression*
    - *Neural networks*
    - *Ensemble methods*
    - *Social network metrics*
    - *Bipartite graphs*
    - *Community mining*
    - *Visual analytics*
    - *Model monitoring and backtesting*
* *Why is my book different or better in timing, content, coverage, or approach?*
  + *The above book presents the fraud analytics at conceptual level. Data scientists who need hands-on code examples cannot benefit much.*
  + *In contrast, my book covers state-of-the-art machine learning techniques as well as code examples for data scientists. This also is the reason why the author believes it is a book for the Oreilly book.*

## Book Outline:

*Until we can envision your book in exceedingly concrete terms, we cannot know whether to sign it.*

## Specs and Schedule:

*How many pages do you expect the book to be?*

*400 pages*

*What use you will make of illustrations or screenshots?*

*What special considerations apply to your plans for the book, including unusual format, use of color, hard-to-get illustrations, or anything else calling for unusual resources?*

*The draft version of this book is written in Jupyter Notebook. It is compatible with O’Reilly’s Atlas platform for publication.*

[*https://www.oreilly.com/ideas/jupyter-at-oreilly*](https://www.oreilly.com/ideas/jupyter-at-oreilly)

*When do you anticipate delivering a complete draft of the manuscript or technical review?*

*February, 2018*

**Complimentary content:**

*What additional topics or content (existing or not) would appeal to your audience?*

* *Real-time fraud detection system*
* *Introduction to a real-time fraud detection system with AWS*

*What additional supplementary or complimentary content would you be willing to coordinate with us to complete? (blog posts, webcasts, video course, online training)*

* *Blog posts*
* *Online training*