**QINHUI XU**

(202) 999-6163 • 1600 S Eads St • Arlington, VA 22202 • qinhuixu0104@gwu.edu

[www.linkedin.com/in/qinhui-xu-a37941a0](http://www.linkedin.com/in/qinhui-xu-a37941a0) • <https://www.kaggle.com/qinhui0104>

SUMMARY

Kaggler/ Data Scientist with a master degree in Business Analytics and hands-on experience in statistical modeling, quantitative analytics, data mining and machine learning. Highly skilled in building explainable machine learning models and creating interactive plots and shiny applications to help make business decisions.

* Extensive experience using Pandas, H2O, Scikit-Learn, Spark, TensorFlow, NLTK, etc.
* Proficient in Python, R, SQL, Tableau, Scala, Matlab.
* Kaggle Competitions Expert: One Silver Medal (36th/3274), One Bronze Medal (147th/2488)

EXPERIENCE

**RiskSpan Arlington,VA**

***Quantitative Model Analyst/ Financial Data Consultant*** *May 2018- Now*

* Built R Shiny application for model sensitivity testing which enables users to easily transfer files between local machine and remote server, update request xml file and set up model run with different config files based on ssh package. Realized functions to parse model result into readable format, generate variable sensitivity report in pdf format which can be directly downloaded into local machine or sent to email address as attachments using ggplot2, openxls and Outlook API. Users can also easily do model result comparison with this application.
* The application can **reduce 95% of the manual work** for sensitivity testing each time new version of model released, which can **save at least 3 hours** for one testing case.
* Built Generalized Linear Model to predict short term mortgage rate using H2O after feature creation based on seasonality decomposition result and business cycle definition. Used K-Fold cross validation to evaluate and select the best model, where **R square is 0.963**.
* Fed the mortgage rate model result into prepayment model and compared the predictions with actual prepayment and previous predictions. **The accuracy of prepayment model got improved**.
* Participated in building scalable Machine Learning Models including Random Forest Model, Gradient Bossting Machine and Naïve Bayes to identify the most possible matched column name based on existing column name and column data. Used Grid Search for Hyperparameter tuning. Compared the model performance based on ROC Curve. **Model can help ensure data consistency and save more than 30% of the data cleaning work.**
* Created Multiple Regression Models with Level, Volatility, Reversion and Momentum Factors running Monte Carlo Simulations for scenarios analysis – National HPI, Unemployment Rate and Mortgage Rate. Did state level HPI projections based on National HPI simulations using Multiple Regression with error correction term using Scipy. Divided simulation paths into three scenarios – optimistic, adverse and severely adverse scenarios based on client’s requirements and fed three scenarios into RiskSpan CECL module with calculated weights for stress testing and more accurate loss prediction.
* Provided consulting service at Fannie Mae on MultiFamily Accelerate the Close Project. Automated shocked input files generating process using Python. Did sensitivity testing on 10Y Treasury Rate, Unemployment Rate, DSCR, LTV and Delinquency Rate at book level. Built attribution dashboard to visualize the difference in Gross Loss, Net Loss and Sensitivity for Acquisition loans and Non-Acquisition Loans for senior level management team using Tableau.
* Did back testing based on historical results using Python. Generated factor grid using Python and assisted in developing SAS script for generating Next Month Loss Estimate based on user inputs.

**TransVoyant Alexandria, VA**

***Junior Data Scientist*** *May 2017- May 2018*

* In charge of data cleaning and data exploration on the flight dataset read from AWS (**4M+**) using SparkSQL.
* Built **ARIMA, Neural Network and Exponential Smoothing models** to predict congestion of airports and average delay time for flights using SparkR. Added dummy variables to ARIMA model which can lower down RMSE. Did Cross Validation to check the accuracy of model. Did model stacking to improve the accuracy. Developed scripts for setting up daily run to update prediction results and push the model run result to front end.
* Built heatmap, barplot, interactive line chart over map with leaflet and flexdashboard in R to provide insights for clients.
* **Minimize RMSE to 0.25**, which can **help clients reduce cost on air transportation by 10%**.

**THE WORLD BANK GROUP Washington, DC**

***Data Visualization Intern*** *Jan 2017 – May 2017*

* Cleaned survey dataset on business practices with 100+ variables using SQL and explored distribution of target variable –annualsale using **Tableau and Python**; Built **3D plot and correlation plot** to visualize the interaction between variables like age and sex using ggplot2 on R; Got cluster characteristics with **K-means clustering method** on H2O.
* Created animation plot with **gganimate** and **ggmap** on R to visualize the changes in trade balance from 2007.

MACHINE LEARNING PROJECTS

**Kaggle Competition: Sberbank Russian Housing Market Washington, DC**

* **Ranked 36 from 3274** (top 2%); Got the Silver Medal; Led two-member team
* Explored data using seaborn and bokeh package on Python. Selected important variables based on variable importance plot from H2O. Feature engineering based on text analytics, sentiment analytics, K-means and Rate by Level method.
* Built Random Forest Model, Generalized Linear Model, Gradient Boosting Model, XGBoost and Neural Network with Sckit-learn, H2O, XGBoost package using Python. Did grid search for parameter tuning. Used K-fold cross validation to compare and select best model. Built three-level model stacking using Stacknet.

**Kaggle Learning Competition: Credit Card Fraud Detection Washington, DC**

* https://github.com/Qinhui-Xu/Credit-Card-Fraud-Detection
* Did data exploration using seaborn and pandas with Python. Visualized data to research outliers using Tsne.
* Dealed with imbalanced data with different resampling methods to improve Recall rate of prediction result. Built Logistic Regression Model, Random Forest Model, Gradient Boosting Machine, XGBoost and Neural Network models on both of undersampled dataset and oversampled dataset using H2O and Sklearn. Tuned parameters of different models using Grid Search.
* Reduced dimensionality of original dataset using Autoencoder and then built different models based on new dataset.
* Compared ROC Curve and Recall Rate of model results. Highest Recall rate is 0.999965 using Random Forest Model on an oversampled dataset based on SMOTE method.

**Analysis and Visualization of the Request Service Number in Austin Washington, DC**

* <https://github.com/Qinhui-Xu/istm-6212/blob/master/Project03/Project03_Final.ipynb>
* Cleaned data of 38K+ transactions using Spark. Designed database and realized star schema using SQL. Did nested queries using SQL with defined foreign key and primary key. Realized interactive heat map, line chart and bar chart using **Plotly** and matplotlib on Python based on result of queries.
* Introduced spending dataset and analyzed relationship between it and service requests based on linear regression.

VISUALIZATION PROJECTS

**Kaggle: Suicide Dataset Exploration Washington, DC**

* <https://qinhuixu0104.shinyapps.io/Kaggle-SuicidesDatasetExplore/>
* Did data cleaning using SQL in R. Built an interactive R shiny dashboard including line chart, pie chart and map.
* Realized interaction between charts so users can see the difference in suicides distribution between different countries, ages, sexes and year.

**Analysis and Visualization of the Social Network of American Football Washington, DC**

* <https://qinhuixu0104.shinyapps.io/FootballNetworkVisualization/>
* Visualized the whole network, subset network and community plot by Shiny using R based on igraph and ggplot2 to show the difference between different universities and between conferences.

EDUCATION

**THE GEORGE WASHINGTON UNIVERSITY, Master of Science in Business Analytics,** GPA: 3.80/4.0*Dec 2017*

* Areas of focus: Data Mining, Statistics, Big Data, Machine Learning, Data Visualization, Time Series

**EAST CHINA NORMAL UNIVERSITY (ECNU),** **Bachelor of Engineering in Software Engineering,** GPA: 3.35/4.0 *Jun 2016*

ADDITIONAL INFORMATION

**Technical Skills:** Python, Scikit-Learn, TensorFlow, R, SAS, Spark, AWS, D3.js, SQL, Linux, VBA

**Certificate:** SAS Graduate Joint Certificate in Business Analytics