Project Descriptions

Virtual Electricity Futures Trading

[*This is currently in development and has not been deployed*]. In collaboration with a large Midwestern electric services company, a virtual electric futures energy trading system is being developed. This serves as a for-profit trading system similar to equity and options trading. The data to serve this project include several daily and intra-daily data feeds. Data are pulled via multiple API endpoints as well as FTP storage locations, and generally fall into three main buckets. The first bucket is weather data, where hourly temperature, wind, rain, etc. forecasts are being scraped at several locations across the United States and Canada. The second bucket is market operational data, including load, wind, and solar forecasts, as well as day ahead electricity price predictions across multiple "nodes" spanning the United States. The third bucket is fuel data, where current and forecasted gas price (and related metrics) data are queried. The data from these three buckets are centrally stored in an internal database. Machine learning models are currently being developed to predict day ahead prices (which is the operational trading unit) simultaneously across every node in each market and at every hour of the day. Multi-input, multi-output neural networks are developed that integrates all the aforementioned data sources and generates a matrix of predictions. The next goal of this system involves collaboration into post-processing trading strategies (including the equivalent of short and long positions, along with anomalous price spike and outage predictions). Historical back testing, along with validation testing in a small market has shown promise for a profitable trading system.

Lead Scoring

Developed lead scoring neural networks to serve as a decision support tool for insurance lead generators, aggregators, and insurance lead buyers to help screen out low quality insurance leads. Historical dispositions data were extracted across multiple lead buyers and were combined in a central data store and staged for analysis. Neural networks were trained and were tuned to provide multiple predictions surrounding lead quality, including the probability that a given lead will: be contacted/reachable, be flagged as a bad lead or otherwise returned, and will convert or end up in an eventual successful sale or transfer. New disposition data are automatically ingested via external database queries and flat-file SFTP transfers, and neural networks are retrained once per month. This scoring model is currently used in production in a "PING-POST" system. Here, an API was developed to receive "pings", or partial lead information, from several distinct lead publishers. Each ping is scored in real-time using the persisted neural network. If scores exceed a given threshold, then those pings are then sent to a network of buyers. The API then collects multiple buyer bids and passes along the highest bid back to the original lead publisher destination. Depending on if a ping was accepted by the publisher, the API then receives a "post", or the full lead information including PII, which is subsequently transferred over to the lead buyer. This model currently produces more than 2 million individual predictions every day and is used to help screen out low-intent pings/leads for insurance lead aggregators and buyers.

Margin Scoring

Developed a margin optimization neural network to serve as a decision support tool for insurance lead aggregators. Historical dispositions data were extracted across multiple lead buyers and were combined in a central data store and staged for analysis. Data include dispositions (e.g., did a lead convert into a sale), along with price/bid information, sale/transfer information, and historical margin proportions. 12 separate neural networks were developed to serve across different data types and industry verticals. Multi-input, multi-output models were trained to predict the win/sale probability, the optimal predicted margin to take on a lead (along with quantile/distribution predictions), as well as disposition predictions like the "Lead Scoring" use case above. These NN models are persisted and hosted in a real-time API that receives hundreds of thousands of requests every day, of which predictions are passed back in real-time. Based on newly scraped and ingested lead data, these neural networks currently are set to retrain once per week. The motivation for this model centers on lead aggregators needing to supply competitive price bids to lead publishers and associated margins taken from lead buyer supplied bids. The client has seen win rates increase by 50%, and profit per lead averages increase by 25%.

Smart Watch Predictive Monitoring

Developed new neural networks and optimized data pipelines for a smart watch predictive monitoring solution, which is currently used in a cancer patient population in a large Midwestern hospital. Data are scraped in near real-

time from Garmin's API, which brings in metrics related to physiological health (e.g., sleep data, stress indices, anomalous event data). These data are merged with clinical EMR/hospital data. Using clinical events data, models are trained using the smart watch data to predict adverse event probabilities. These probabilities include the chances of unplanned hospital admissions, emergency room visits, appointments, and other fatal event predictions. Predictions and outputs are then shared with hospital key staff and physicians through 1) flat-file prediction files, 2) customized HL7 messaging that integrates with the hospital's EMR system, and 3) an interactive dashboard.

Propensity to Pay Healthcare

Created data pipelines to ingest nightly EMR billing data dumps via scans across a maintained SFTP server. The data include personalized details surrounding statements, charges, adjustments, payments, and demographics data. Ensemble machine learning based propensity to pay scoring models were trained using this data, and predictions are made on a patient's 1) likeliness to repay on an outstanding collection, and 2) the best outreach medium (e.g., SMS, phone, or email). These models are currently used in production across 25 individual hospitals, have increased collections by 20%, and have helped to decrease outreach volume through personalized scoring, which simultaneously helps mitigate the "over-contacting" problem present in call center environments. These models are persisted and re-trained daily, and predictions are sent to the client via flat-file transfers, as well as through a customized interactive dashboard to link scoring models to outcome data (e.g., payment collections).

Email Marketing Scoring

Developed neural networks to power an email marketing solution that predicts several quantities of interest for marketing emails (open, click through, opt-out, and sale generation probabilities). This neural network is deployed in an API that receives hundreds of thousands of lead requests per day, which are parsed and passed through the predictive ML model, where multiple correlated outcomes are backed back from the API. The data to serve the model include historical touchpoint attempts, and metrics related to a user's interaction with the email. These data are provided in nightly data dumps, and the model is re-trained every month.

Biopharma Testing and Revenue Forecasting

Created data pipelines to ingest nightly data dumps via SFTP into a centralized data store. Data include historical testing volume, revenue data, and contract details from a biopharma testing company. From these data, multiple timeseries forecasting models were trained to predict short term (6 week) and long term (12 month) revenue and testing volume. These models are currently in production and are re-trained once per week. A simulation tool was also created that uses historical data to simulate the course of a clinical trial, including enrollment trends, attrition rates, testing schedules, and various test price points. A customized client-facing interactive dashboard was created to house predictions, show historical trends and patterns, as well as an interface for the testing simulation tool. This is currently used by the client as a financial aid and staffing optimization decision support tool.

Cryptocurrency Scoring

Using historical cryptocurrency price data, volume data, and other on- and off-blockchain data sources, a viability scoring model was developed. Using a combination of statistical models including time series principal component analysis as well as representational similarity analyses, this tool clusters tokens to predict its long-term viability across its early developmental years as a cryptocurrency project. Predictions are stop-light coded for easy interpretation. This model is currently in a beta-version, with plans to incorporate automated scoring to be used and consumed on an internally hosted executive dashboard.

Cryptocurrency Pattern Matching

Inspired by traditional technical analysis of financial data, a pattern matching/scanner tool was created. Here, tools and methodology were developed to take images of known technical analysis patterns, and automatically extract and normalize the numerical data serving that pattern. These patterns are then used and compared against current and historical patterns across a large number of cryptocurrency projects, and can be used across any project and across any time scale (e.g., 10 minutes, 2 weeks, 10 years). This tool rapidly scans across multiple projects, and continually

updates and ranks best matching patterns. This is currently deployed in a beta version that is used as a decision support tool within an internally hosted executive dashboard.