Google Analytics Reference Pattern
Investment Products Recommendation Engine
Advanced Analytics White Paper

Oleksii Lialka

Staff Data Scientist, SoftServe Inc.
olial@softserveinc.com

June 16, 2021

Abstract

This document provides an overview of the advanced analytics services and solutions of the *Investment Products Recommendation Engine* (IPRE) reference pattern. It contains sections dedicated to data pipelines and ETL, advanced analytics and machine learning methodology, principles of the recommendation service. The GitHub public repository hosts the source code of the IPRE reference pattern.

1 Solution overview

The IPRE is a comprehensive investment banking solution, which builds a bridge between retail investors and the complexity of the capital markets. The AI/ML products assist clients in optimal portfolio construction, wealth allocation, and discovering key performance metrics of the investment. An ensemble of ML solutions produces highly personalized recommendations and investment advice based on an individuals' risk preferences and wealth management goals.

The reference pattern requires minimal manual effort to set up and deploy advanced analytics services. *Terraform* script builds and deploys all components of the IPRE advanced analytics.

1.1 Project structure

The source code of the advanced analytics services is nested under the `advanced-analytics/` directory. Scripts content follows a modular structure according to the objectives of ETL, portfolio optimization, investment analytics services.

The `data-pipelines/` directory contains Cloud Functions scripts for generating datasets, writing data from Cloud Storage to BigQuery. Modules for *capital markets* and *investor risk preferences* datasets have dedicated Cloud Functions for invoking training and inference jobs of BigQuery AutoML and ARIMA models. Directory `recommendation-engine/` provides scripts for convex optimization, investment analytics.

2 Data pipelines

The IPRE service relies on multiple data sources, both internal and external. The solution implements scalable data pipelines with the Big Data technologies, such as *BigQuery, Cloud Storage, Dataflow*. Data pipelines follow the schema-on-read data lake convention.

The Cloud Functions trigger dedicated scripts to extract, preprocess and write raw data streams into Cloud Storage buckets. Writing an object to the

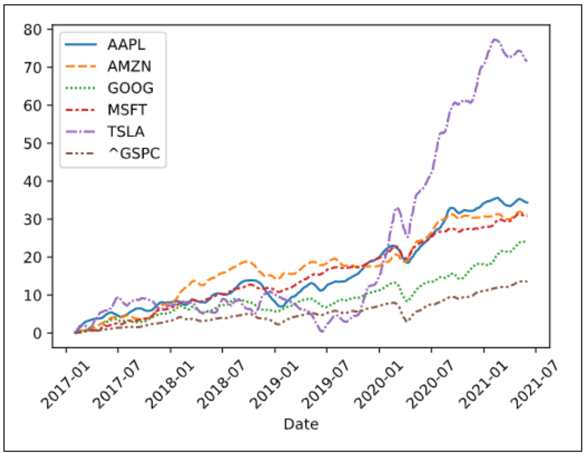

Figure 1: Cumulative monthly returns of sample stocks.

Cloud Storage bucket triggers the Dataflow job for adding new data to BigQuery.

Such architecture makes an ETL pipeline resilient to corrupt data and scalable to multiple data sources. Dataflow provides a clean, cost-effective solution for migrating massive datasets from the data lake to DWH.

2.1 Capital markets data

Historical market data is a crucial element for the recommendation service. A dedicated data pipeline job collects quotes of the select securities from *YahooFinance*[1]. Securities of interest include the Big Tech stocks, FX, ETF, and others. All select assets vary in return and risk. It allows IPRE to construct a wide range of portfolios to meet diverse investors' preferences.

Data preprocessing script removes corrupt and missing records and transforms daily historical quotes (q) into *periodic returns.*

$$r_{t,t+dt} = \frac{q_{t+dt} - q_t}{q_t} \; : \; t \geq 0, \, dt > 0 \qquad (1)$$

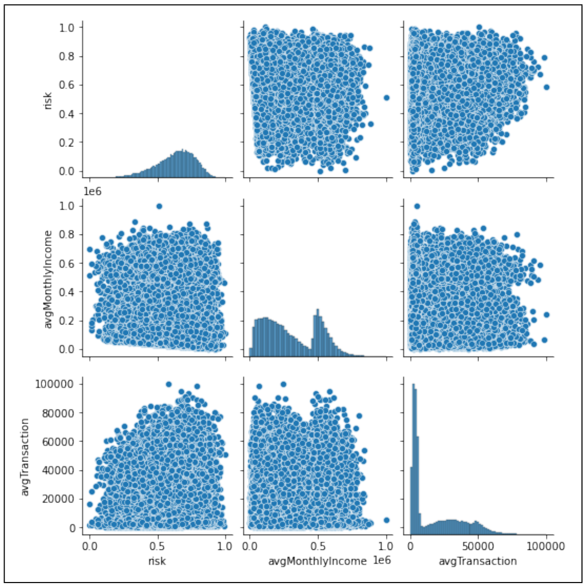

Figure 2: Pair plot of investor risk preferences features subset.

The script writes observations of *returns* with a unique timestamp to the Cloud Storage bucket. It allows to reduce egress and ensures that BigQuery does not receive duplicate data. During the first run of the script, all observations starting from 2017 will make it to BigQuery. Subsequent runs provide incremental observations of the "unseen" data.

In the final stage of ETL, the Dataflow job writes processed data to BigQuery. Aggregating data in BigQuery allows other services to retrieve the data in a cost-effective way.

2.2 Investors risk preferences

The investor risk preferences (IRP) is a synthetic dataset containing historical records of 1000 existing retail investors[2]. This dataset is a crucial component for making personalized recommendations based on an individual's investment preferences.

Risk aversion is a target variable of interest. *Av-*

[1] Modular interface of data pipelines permits integration of any other data provider

[2] The dataset mimics customer data FSI companies typically collect. The solution can handle real clients' data.

erage monthly income, education, loans, deposits are among 15 independent variables. A set of unique continuous variable distribution functions such as *Gamma, Gumbel, Gaussian*, etc., generate investors' attributes. A script produces monthly snapshots of investors' attributes, resulting in 48'000 data points.

The Cloud Function triggers a generation of the dataset upon the first launch of IPRE. Dataflow migrates generated dataset from Cloud Storage to BigQuery.

3 Advanced analytics

This section describes the use of BigQuery built-in tools for ML and forecasting. The BigQuery AutoML service produces factors for the recommendation engine. End-users can use BigQuery ML and forecasting tools to leverage massive datasets to produce deep customer insights and business analytics.

3.1 Risk aversion

Predicting the risk aversion of prospect investors is a cornerstone IPRE feature. The risk aversion factor is a proxy of an investors' risk preference. Individuals' risk preference corresponds to the level of risk an investor sustains within the long term. Modern financial theory and behavioral finance studies advocate that the risk aversion factor shapes the investment decisions of an individual[3].

Inferring the risk aversion from individuals' attributes is a supervised machine learning problem. The Cloud Function invokes a script for training the BigQuery AutoML model on the IRP data stored in BigQuery. BigQuery ML tools handle the entire model training pipeline, from data preparation to hyperparameter tuning.

Once training is complete, an inference script predicts risk aversion for the set of existing clients who are not yet active investors. The mean objective of the IPRE is to provide non-finance professionals with an insight into individual's risk preferences.

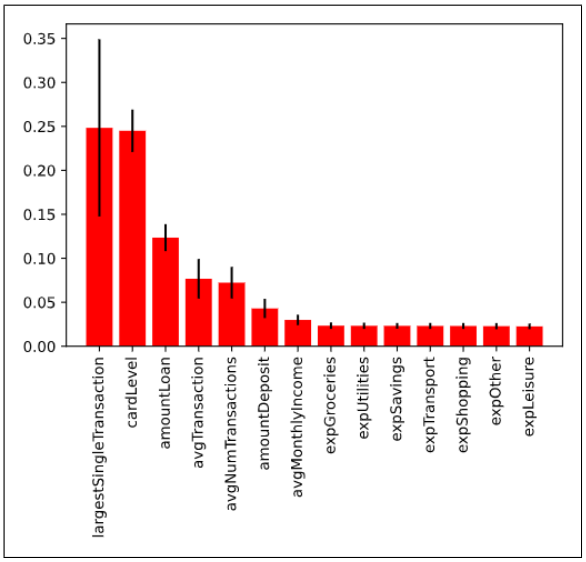

Figure 3: Feature importance of the investor risk aversion model.

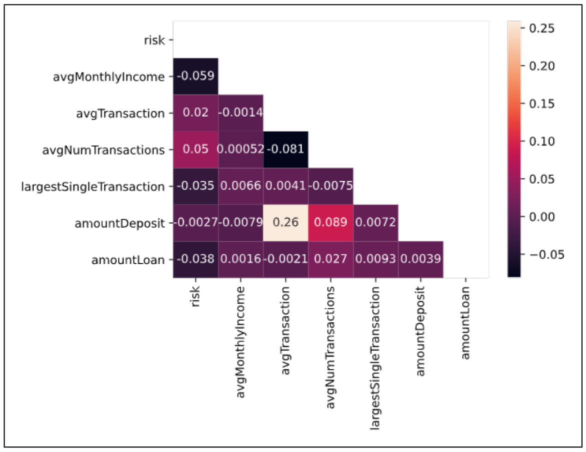

Figure 4: Correlation map of investor risk preferences features subset.

[3] Björk, T., Murgoci, A. and Zhou, X.Y. (2014). Mathematical Finance, 24: 1-24.

3.2 Expected returns

The IPRE utilizes portfolio optimization techniques for producing a recommendation. The expected returns vector (μ) is a predicted return on standalone assets (stocks, indices) for an arbitrary investment horizon. The precision of the recommendation depends on the accuracy of expected returns.

Processed historical market quotes turned into periodic returns and stored in BigQuery. The Cloud Function triggers the BigQuery forecasting tool to train models for each time series stream. Invocation of multi-threaded BigQuery jobs reduces overall training time. BigQuery uses an advanced version of ARIMA for time series forecasting. The BigQuery implementation of ARIMA is a good fit for forecasting noisy, non-stationary financial time series data.

Models make a one-step-ahead prediction of the expected returns and store results in a dedicated Cloud Storage bucket.

4 Recommendation service

The IPRE service takes user ID as an argument, retrieves predicted risk aversion factor, and makes optimal portfolio recommendation. The risk aversion factor is inferred from the clients' data using the BigQuery AutoML model. The user can select a preferred level of risk with the UI to obtain a recommendation, which fits them best. The IPRE returns a recommendation on the capital market assets, expected risk, return, portfolio performance metrics.

4.1 Portfolio optimization

Maximizing the difference between expected return (μ) and a risk (Σ) by computing the unique combination of assets is the objective of portfolio optimization. The risk aversion factor (λ) adjusts asset weights according to an individual's risk preferences by introducing a penalty to an objective function[4].

$$\max\ \mu^T\omega - \frac{\lambda}{2}\omega^T\Sigma\omega\ :\ \forall\,\omega \in \Re \qquad (2)$$

The higher the risk aversion, the higher the penalty for including risky assets into a portfolio. Portfolio optimization produces optimal asset weights in the portfolio given the risk aversion level. Users can explore different portfolio compositions by adjusting a risk preference factor in the UI.

4.2 Investment analytics

The IPRE service calculates a set of portfolio key performance metrics: expected return, risk, and *the Sharpe Ratio.* The investment analytics service assists users in making an informed decision on the investment, which fits their risk appetite and investment goals.

5 Implications

The solution to the first-order condition of equation (2) provides the ground truth of the risk aversion factor for *existing* investors. The IPRE can scale to produce recommendations on the OTC assets and structured investment products offered by financial companies to retail investors. Compared to an inhouse solution, the BigQuery tools for advanced analytics, ML, and forecasting allow leveraging massive datasets with lower manual effort and maintenance overhead.

The IPRE unlocks a new dimension of *proactive* recommendation engagement. Serving professional investment advice via web and mobile interface is a financial services industry differentiator. It provides a competitive advantage to a company seeking to increase retail investors' conversion and onboarding rates. Providing users with content-aware, highly personalized recommendations increases the trading activity of existing and prospective retail investors.

[4] MATHEWS, T. (2004). Portfolio Selection with Quadratic Utility Revisited. The Geneva Papers on Risk and Insurance Theory, 29(2), 137-144.