Problems submitted to the 11th Montreal IPSW

Air Canada
Relationship and lags between search and booking trends

Online travel searches can be considered as intention to travel and are often a necessary first step to the actual booking action. Surges in searches can sometimes indicate a sudden increase in the passenger demand, while an increase in the "look-to-book" rate can unveil lack of competitiveness or website issues which present many opportunities for the Revenue Management analyst and can help improve the purchases and booking experience.

By using booking and search data, if Air Canada could identify the relationship between the two signals and build an outlier detection model that would identify changes in those patterns, then they would be able to correct their forecast or modify their inventory policy.

National Bank of Canada
Data anonymization

Data security and confidentiality are at the heart of organizations' concerns today. The challenge is to find a compromise between valuing data and the risk of using personal data for analytical and artificial intelligence projects.

Data governance shares the responsibility with other teams, namely with cybersecurity experts, to put in place the necessary measures to maintain granular control over the access rights of each employee. For example, it defines access levels by role directly on the data. However, the implementation of these measures and the resulting processes restrict access to the data, regardless of what you want to do with it. This approach also reaches its limits when it comes to giving partial access to databases, as the risks of inference and re-identification increase.

Alternatives exist and constitute very active areas of research. They can be grouped into three categories (not separate):
(a) Data anonymization techniques (eg "k-anonymization") aimed at making the original data less sensitive;
(b) Analysis designed to keep data confidential (e.g. "Differential Privacy"). We use the original data and we desensitize the calculations on the data;
(c) Generative synthetic data models.

These approaches introduce the concept of confidentiality measurement and therefore a quantity associated with data risk. Given their inherent advantages and disadvantages, what approach(es) should be favored in order to optimize the compromise between the risk of confidentiality and the valuation of data?

Coveo
Analysis of eCommerce clickstream data for intent prediction and recommendations

Coveo is sharing a rich anonymized dataset on eCommerce behavior for research purposes. The dataset includes more than 30M single product interactions by real shoppers – the clickstream data is enriched by (vectorized) catalog meta-data and by fine-grained search behavior, showing not only products clicked after a query is issued, but also products seen and not clicked (i.e. negative feedback).

In continuity with previous research (from Coveo and the community) and with the 2021 SIGIR Data Challenge, we welcome research addressing two challenges:
(a) a session-based recommendation task, where a model is asked to predict the next interactions between shoppers and products, based on the previous product interactions and search queries within a session;
(b) a cart-abandonment task, where, given a session containing an add-to-cart event for a product X, a model is asked to predict whether the shopper will buy X or not in that session.

Participants are encouraged to read the following peer-reviewed articles: https://arxiv.org/abs/1907.00400 and https://rdcu.be/b8oqN, to get a better understanding of the underlying theoretical and practical issues: class imbalance, conversion rate, etc. The SIGIR Data Challenge paper contains extensive literature review, in-depth details about the dataset, and discussions about the target use cases; the SIGIR Leaderboard can be consulted for quantitative benchmarks (test files to replicate the submission phase can be requested to Coveo); the SIGIR open source repository contains detailed explanation on the dataset, ready-to-use baseline models and links to successful implementations.

Environment and Climate Change Canada
Development of a weather text generator

Weather forecasts include many ways to express possible forecasts. However, the texts of the forecasts are very structured and very limited in their formulation. In

addition, there is only one “good” standard way of reporting the weather considering a given set of concepts.

In order to continue to provide quality weather forecast and information services to Canadians, the MSC wishes to develop a weather forecast text generator, in English and French, which uses meteorological concepts representing, in coded form, the weather forecast.

IATA
Airline Codeshare correlative analysis

The IOSA Audit serves as an industry standard that is used among cooperating airlines. A frequent mode of cooperation is code-sharing. IATA would like to discover and understand the IOSA market and its needs through the lenses of code-share relationships, using traffic data.

The objective is to understand where, to which width and depth airlines codeshare with each other, and what is the correlation to IOSA registration.

Incidentally, IATA would like to explore any other potential correlations with the outputs produced by the airline, for example, explore the relationship between codeshare activities and crucial airline metrics.

Société générale
Categorical variable selection in risk modelling

In financial institutions, categorical features appear quite often in credit datasets and in compliance models, for example, features related to clients’ risk profile.

Traditional feature selection methods (e.g. statistical significance, recursive feature elimination, LASSO) do not work well with categorical features since these methods would retain certain levels and remove other levels of the same feature. The Group Lasso approach has shown to be more stable in terms of variable selection but displays shortcomings in terms of predictability. Instead, for a given feature, would it be more appropriate to devise a method that aggregates neighbouring levels in bins in order to get a feature representation space that would better scale with the output?

Because of the numerous ways to represent categorical variables and to select which variables are of importance, we ask what are the most appropriate methods for improving categorical feature selection.

TD Canada Trust
Portfolio risk forecast

In practice, portfolio managers use either the fundamental model or the statistical model, or a combination of both, to model the volatility of a portfolio's future performance.

The fundamental model specifies characteristics that influence the covariances of equity returns. For example, stocks in the same industry will have more correlated returns with each other than stocks in different industries. Financial leverage or the size of firms also seem to explain part of the common variation in returns. Using a list of these characteristics and their time series as well as historical stock returns, it is possible to estimate the covariance matrix of risk factor returns in order to calculate a security's exposure to these factors. In addition, by aggregating them, we are able to calculate the risk of any portfolio.

The statistical model starts from a full covariance matrix. This has often been calculated with returns filtered by a "winsorization" process. A principal component analysis (PCA) identifies the main dimensions of the covariance matrix by ranking them by their importance but without identifying the factors to which they refer. The risk factors are inferred empirically.

Until now GPTD has built risk models according to these two types of modeling. Statistical models (which use less information) work well, especially when stock markets are less volatile but are less adequate when market volatility increases sharply. So far little consensus exists on the optimal way to combine the fundamental model and the statistical model or on the possibility of constructing a third one. TDAM is looking at what would be the best way to forecast the risk of a portfolio that uses historical returns on stocks in a certain universe and the fundamentals of those stocks.