Background:

HSBC was looking to leverage AI to analyze large volumes of alternative data like news, social and earnings transcripts to identify high-growth stocks and build an intelligent investment portfolio. They wanted to make advanced AI investing strategies used by institutions accessible to retail investors. However, the amount of relevant data was overwhelming.

HSBC partnered with EquBot, an AI specialist that utilizes IBM Watson's NLP and machine learning to power its AI investment platform that evaluates companies.

Together, they created the HSBC AI Powered US Equity Index (AiPEX) which is an AI-selected portfolio of around 250 stocks with strong growth potential based on financials, news sentiment, and management quality. It outperformed the market by over 120% in the last decade by detecting signals in textual data that humans missed. AiPEX recognized investment themes around industries like energy and pharmaceuticals earlier than traditional indexes.

The core of EquBot's platform is powered by Watson Discovery, which analyzes millions of news articles daily in multiple languages, and Watson Studio, which ensures model accuracy. The IBM partnership lends AiPEX credibility. By explaining the AI selections in an interpretable way, HSBC has built trust and strong interest in AiPEX. Client events at IBM experience centers further showcase AI capabilities. The success of AiPEX underscores the significant disruption AI and alternative data can drive in investment management. Building on momentum, HSBC plans to leverage AI to develop more specialized investment portfolios aligned to emerging themes detected in the data.

This project thus provides an ideal testbed for my own research into AI and machine learning innovations in finance.

Response:

Introduction

This report provides recommendations for implementing natural language processing, time series forecasting (ARIMA and Exponential smoothing), random forest, ensemble, and portfolio optimization models to power the AI-based stock selection and trading platform as would have been modeled through the partnership of HSBC and IBM. Integrating these techniques will allow robust analysis of alternative data to identify high-potential stocks and optimize the risk-return profile of the portfolio.

"To select the stocks that make up AiPEX, the EquBot AI investment platform follows a rules-based investment approach and monitors the businesses listed within the Russell 1000 index, identifying the roughly 250 companies with the highest growth potential according to its calculations. "For the financial score," explains Robertson, "it's looking at things like cost of goods sold, price to earnings ratio, revenue expenses, earnings per share. And the second score, the news and sentiment ranking, is determined by Watson as it reads the news articles and social media to see what people are saying about specific companies. Finally, the management score looks at the actions and performance of management and the C-suite of these companies as well as how they are perceived in the market."" https://www.ibm.com/case-studies/hsbc-usa

Model #1: Natural Language Processing

Given **earnings call transcripts, financial filings, news articles and sell-side investment analyses for public companies across sectors,**
Use **natural language processing to extract signals about growth drivers, management sentiment and business strategy.**
To **identify undervalued stocks with improving growth potential ahead of human analysts.**

I would imagine that the IBM Watson NLP model can encode unstructured textual data into vector representations and extract relevant features like entities, semantics, and sentiment scores. Fine tuning large pre-trained models on financial corpora would enable context-aware analysis of news, earnings calls and SEC filings to detect growth drivers for companies. The models would consume thousands of articles daily across news and social media to stay updated.

Pre-trained NLP models that have learned general linguistic representations can be fine-tuned on financial corpora like earnings call transcripts, financial reports, industry news articles etc. This teaches the model the nuanced language used in finance to understand the specific semantics. For example, a fine-tuned NLP model can classify text snippets from earnings calls to detect mentions of growth drivers. Sentences discussing expansion to new markets, positive competitive dynamics, product innovation and commercialization events could be classified as references to growth. These provide signals used in stock selection. The fine-tuned models can also extract sentiment scores on management outlook commentary in transcripts. A more positive outlook language may correspond to higher estimated growth potential. Time series modeling of sentiment can then help to uncover improving/deteriorating management optimism.

For example, Microsoft's earnings call mentions growth of its cloud platform Azure many times. The fine-tuned NLP model can classify sentences referencing Azure's sales growth, new datacenter expansion, big client wins etc. as strong signals of Microsoft's future growth prospects. On the other hand, if Apple's earnings call sentiment appears overly cautious about iPhone 14 demand amid inflation, the model can score management outlook as more negative vs positive historically. This lower future optimism score flags near-term growth headwinds.

In essence, fine tuning teaches NLP models the nuanced language of finance to uncover forward-looking growth signals. The alternative data in speech and text contains insights that even expert analysts can miss given the volume of data. These NLP models act as an army of virtual research assistants detecting growth drivers. Their predictions help select stocks likely to outperform markets through product, services and market share expansion.

Model #2&3: Time Series Forecasting: ARIMA & Exponential Smoothing

- ARIMA:

Given **over 10 years of historical quarterly financial statement data including revenues, costs, debt levels and other financial indicators for public companies,**
Use **ARIMA (AutoRegressive Integrated Moving Average) time series models that rely solely on past values of each univariate time series to predict its future values, tuned specifically on each company's historical financial statement item data.**

***To* forecast trajectories for quarterly earnings, fundamentals and growth metrics, leveraging the momentum and long-term persistence in each series while adjusting for seasonality, outliers and volatility.**

Many fast-growing tech companies like Tesla show consistent rapid revenue growth over time. Historical quarterly sales data can reveal persistent growth patterns. We can apply time series analysis to identify stocks maintaining high sales momentum. The ARIMA statistical model forecasts future values in a time series based only on its own past values.

For example, Tesla's quarterly vehicle deliveries have increased 30-60% year-over-year the past 3 years. An ARIMA model trained on 10 years of quarterly delivery figures could reliably predict deliveries next quarter will grow another 50%.

Seeing high expected growth rates flags such stocks. ARIMA accounts for seasonal dips too - e-commerce sales often slow after holidays but mean-revert upwards. It would correctly forecast the rebound. Parameters in ARIMA control how much past values influence the future. Fine-tuning these on each company's financial time series customizes the growth forecasts. Automation helps rapidly evaluate parameter combinations to optimize growth forecast accuracy.

Besides sales, ARIMA can model revenue, profits, market share and other fundamentals demonstrating steady historical gains persisting into forecasts. Consistent high predicted growth across indicators helps select innovative stocks outpacing peers. The key advantage of ARIMA is it customizes each company by tuning sensitivity to historical values, enabling it to accurately predict future performance for both high growth stocks as well as slower moving companies. By quantifying expected growth, ARIMA-driven analysis provides uniquely customized forward-looking insight to guide selection of stocks with strong fundamentals persistence that beat benchmarks and analyst expectations.

So in summary, ARIMA leverages historical patterns exhibiting consistent growth to anticipate similar future momentum. Tuning the model on each company quantifies sales and fundamental growth profiles, flagging emerging leaders.

- Exponential Smoothing:

***Like ARIMA, given* over 10 years of historical quarterly financial statement data including revenues, costs, debt levels and other financial indicators for public companies Exponential smoothing could also be *used* as a good model *to* apply for forecasting financial time series and selecting high growth stocks.**

The intuition behind exponential smoothing is that more recent values in the historical time series have greater influence in predicting the future. So recent quarters receive higher weightage. This matches typical growth stock behavior. For example, a hot tech stock can accelerate growth dramatically in just 2-3 quarters due to a new product. The last few datapoints better predict near-term momentum.

In exponential smoothing, a smoothing parameter controls how rapidly older data gets downweighed. Values closer to 1 give more emphasis on recent history. The model is also easy to retrain frequently as every new quarter gets added to the series. This keeps growth predictions timely.

Overall, the automated parameter tuning and flexible incremental retraining of exponential smoothing tailors it to detecting corporate growth inflection points early. Like ARIMA, it enjoys customization to each stock's dynamics.

Model#4: Random Forest

Given **historical stock fundamentals data such as valuations, earnings, balance sheet metrics, macroeconomic factors, analyst estimates, and textual data from news, earnings calls and reports for US public companies,**
Use **random forest classification models comprised of decision trees trained on historical features of high growth stocks** ***to*** **predict new companies with strong growth potential.**

The random forest model would be trained on 10+ years of historical data on both rapid growth stocks and more mature/declining companies. Features engineered from fundamentals, macro data, text analytics and technical market signals are fed into decision trees.

Each decision tree makes an independent stock classification by applying yes/no rules learned during training. The random forest model tallies votes from every decision tree to make a final prediction of expected growth profile.

Comparisons against realized future stock returns enables evaluating feature importance and tuning decision thresholds to optimize predictable alpha. This approach sidesteps complex simulations and time series forecasting to instead learn recognizable historical growth patterns automatically. Automated feature engineering and model tuning optimize out-of-sample growth stock picking accuracy.

Model #5: Ensemble Modeling with Portfolio Optimization Quadratic programming optimization

In practice, a heterogeneous ensemble combining the NLP inferences, time series forecasts and random forest model outputs can amplify the strengths of each approach. A weighted average model based on historical performance can aggregate the signals. Stacking additional models like simulation on all outputs improves robustness.

Portfolio Optimization Quadratic programming optimization adjusts stock weights in the portfolio to maximize Sharpe ratio based on risk-return profiles of selected securities and covariance between their forecast returns. This accounts for diversification and the best sizing considerations.

Individual AI models have strengths and weaknesses. An ensemble combines their outputs to improve overall robustness and stability for stock predictions. Examples:

1. An NLP model analyzing management commentary may over-extrapolate temporary sentiment changes. But ARIMA's reliance on historical fundamentals prevents overreacting to transient noise.

2. Random forest models can overfit spurious correlations. But NLP extracts signals with clear causal links to growth like new product releases.

3. ARIMA provides accurate short-term forecasts but lacks insight into long-term strategy shifts. However, NLP applied to earnings call topic modeling uncovers emerging management priorities early.

4. Random forest model performance degrades rapidly on new unseen data compared to time series models directly fitting past trends. However, it adapts better to highly variable growth profiles.

5. NLP sentiment analysis relies on tone and emotive language. But ARIMA simply projects mathematical trends oblivious to subtle linguistic cues about confidence.

6. Macroeconomic changes can negatively impact ARIMA predictions that extrapolate past company-specific patterns. In contrast, random forests factor in external leading indicators.

So why did I choose my models and what other models could I have considered?

To construct an AI pipeline for stock analysis, I considered a few other modeling choices for each stage. For natural language insights, Naive Bayes provides a simplistic approach to text classification, but risks oversimplified assumptions. In forecasting momentum, time series regression methods like linear models quantify historical correlations and trends. But complications arise if macro conditions change. At the same time, autoregressive models provide a non-linear approach robust to regime changes. K-means clustering presents an unsupervised data mining method to group stocks based on features like volatility, valuation and sector. This enables customized forecasting and optimization tailored to cluster-specific dynamics. Potential risk, however, includes incorrect centroid or suboptimal cluster number selection. And finally, for pattern recognition, the original random forest proposal helps avoid overfitting.

By not only considering the models that I suggested but looking at other techniques, I aimed to select a combination that balances accuracy, optimization constraints and interpretability. But contingencies should be established to respond to potential model degradation through retraining, ensembling or replacement.

A heterogeneous ensemble takes growth probability forecasts for a stock from NLP, ARIMA, Exponential smoothing and Random Forest models. It assigns weights to each model based on historical accuracy. The weighted ensemble average mitigates biases like NLP's sensitivity to language nuances, ARIMA's assumption of persistent historical trends and Random Forest's overfitting.

Once stock selections are made, we feed the results into a quadratic optimization algorithm that computes target portfolio weights for each stock. Weights are based on projected returns, volatility and inter-stock covariances. This maximizes portfolio Sharpe ratio given risk tolerances.

For example, suppose 3 stocks (A, B, C) are selected with forecast annual returns (10%, 18%, 22%) and standard deviations (15%, 30%, 40%). The correlation matrix between predictions is input.

The optimizer then computes weights of 55% in B, 15% in C and 30% in A by solving for portfolio with maximum differential return per unit of risk based on forecasts and covariance. Optimized rebalancing helps minimize concentration risks.

Objective Function:
Maximize Portfolio Sharpe Ratio

Decision Variables:
Weights of each stock
Return of each stock
Volatility of each stock
Σ = Covariance Matrix between all Pairs of Stocks
Constraints:
w_a + w_b + w_c = 1 (Weights Sum to 1) 0 ≤ w_i ≤ 1 (Long Only Weights Between 0 and 1)

By solving the quadratic optimization with these inputs, the model provides an optimal asset allocation customized to risk objectives. In effect, the ensemble harmonizes individual model limitations while optimization provides a well weighted basket of stocks that maximize risk-adjusted return among selected stocks.

Final Thoughts & Conclusion

Finance is one area where there is readily available, real-time data available in large volumes. Companies access publicly available data quite easily from the internet or may be able to buy alternative data (data not directly related to the market) from third party companies. While data cleaning, manipulation and access is still necessary, very rarely would you encounter missing market data issues with publicly available financial data. One of the only challenges is determining how to put this data together to create insight and how to build models that can help to improve predictive accuracy given that market data is often very noisy.

By exploring different AI-driven investment strategies, I've learned that combining various advanced methods can really give us an edge in picking stocks and managing portfolios. Using natural language processing helps us understand information from news and other sources, uncovering important things about a company's growth and how its management feels. Time series forecasting, like predicting future trends with ARIMA and Exponential Smoothing, helps us see where a company might be headed based on its past performance. Then, there's the random forest model, which helps spot new companies that could grow a lot based on different important factors. When we put all these methods together in an ensemble, it's like creating a team that works well together, making our predictions stronger. And finally, using portfolio optimization, we can carefully balance our investments to manage risks and aim for better returns. This blend of approaches not only dives deep into different data but provides an idea of IBM and HSBC's approach to using AI and alternative data to derive alpha from the market.