Factor Prediction: Forecasting Risk

Final Presentation

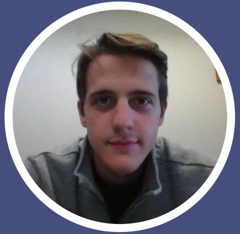
Date 12/6/2018

Iowa State University

Senior Design - sddec18-13

Principal®

Meet the Team



Nathan Hanson

Jack Murphy

Carter Scheve

Caleb Utesch

Samuel Howard

Alex Mortimer

Principal®

Agenda

- Project Introduction
- Project Scope
- Results
- Looking Forward

Agenda

- Project Introduction
- Project Scope
- Results
- Looking Forward

Motivation

Why is this project important?

Current Techniques

Current methods focus on expected return rather than variance or volatility

Quantifying risk is crucial for informed decision making

Solution

Aggregate stock-level data into feature-level data

Utilize Machine Learning and Statistical Modeling

Create software tool for making investment predictions

Advantages

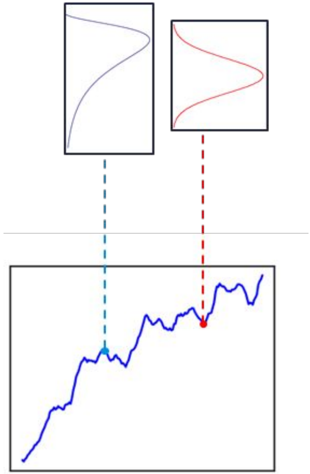
Give portfolio managers better information for their portfolios

Tool can eliminate erroneous human decisions

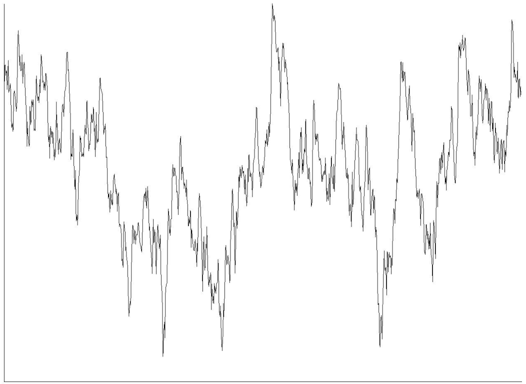
Increase decision-making speed for volatile market

Goal: Forecasting two measures of factor risk

Accurate forecasts of cross-sectional return variance and time-series return volatility would enable better factor selection and support portfolio allocation decisions.



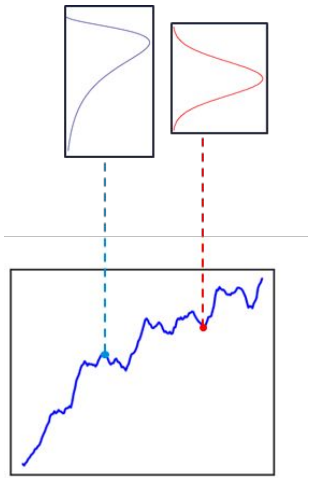
Cross-Sectional Return Variance



Time-Series Return Volatility

Principal®

Goal: Forecasting two measures of factor risk



Cross-sectional Return Variance

A measure of risk that describes the spread of **stock returns** within the factor portfolio over a specified horizon.

Example: Variance of future 6-month returns for the 100 stocks in top decile of Book-to-Price at one cross section of time.

How risky is it to pick a sample from this group?

What if the wrong stocks are selected?

Principal®

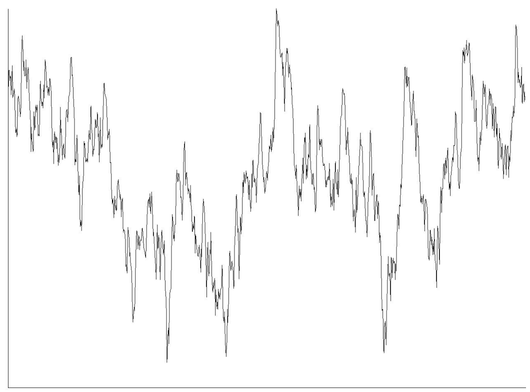
Goal: Forecasting two measures of factor risk

Another measure of risk that quantifies the spread of **portfolio returns** over a future horizon.

Example: Standard deviation of weekly returns of the top decile of Book-to-Price over the future 6-month horizon.

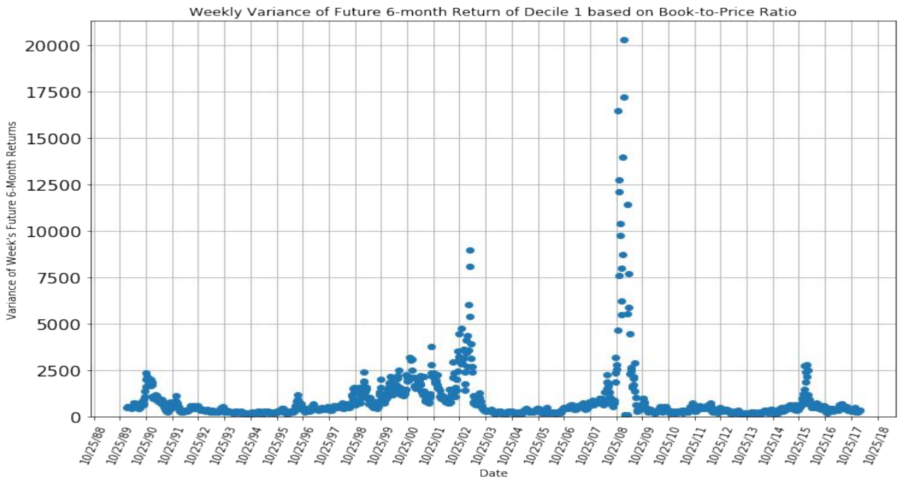
Is this factor likely to generate extreme returns?

Can I tolerate this outcome's uncertainty?

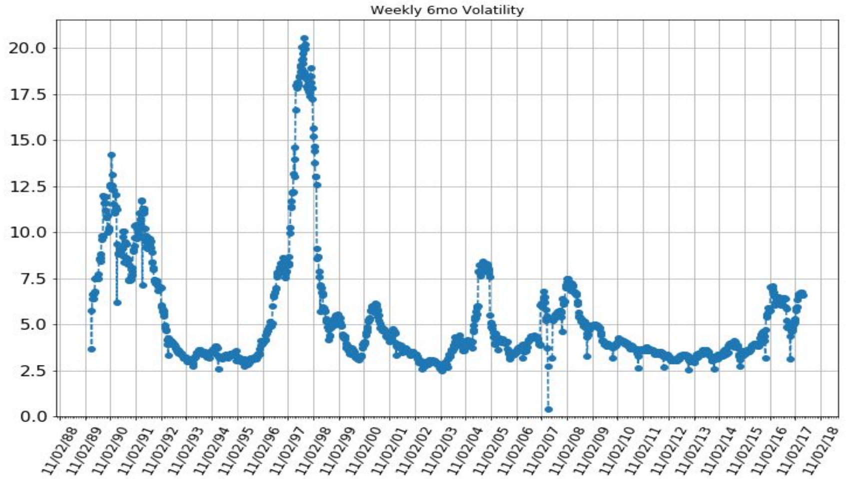


Time-Series Return Volatility

Weekly Risk Visualization



- Left: Cross-sectional variance
- Right: Volatility
- Initial EDA helped to understand the data we are trying to regress
- Transforming the data to a log scale helps



Principal®

Resources

Russell 1000

Amazon EC2 Instance

Anaconda
Jupyter
RStudio

Agenda

- Project Introduction
- **Project Scope**
- Results
- Looking Forward

Principal®

Project Scope

The goal of this project is to explore novel methods for analyzing and predicting certain factors of stocks. We have researched several feature analysis and machine learning techniques to use to generate predictions of future market performance. Once successful, this process could improve investment decisions and reduce the amount of time needed from analysts to manage a portfolio.

1. Research on machine learning and data analysis techniques
2. Organize data as required for models; Create several different types of models to identify prospects
3. Tune model parameters and apply feature analysis to improve results
4. Test effectiveness of models and analysis methods through a historical portfolio analysis

Project Flow

New Idea Deliverables

- Insights and surprises into today's pain points
- Theoretical viable solution
- Organizational support to tackle problem

POC Deliverables

- Preliminary business value defined
- Minimum viable solution
- Organizational, Process and end user assessment

Pilot Deliverables

- Validated business case
- Minimum scalable solution
- Organizational change commitment
- Validated with real data in a real application environment

Enterprise Deliverables

- Fully scalable solution
- Business Implementation
- KPI's in place
- Organizational change
- Process changes implemented

Prototype → Proof of Concept → Pilot → Enterprise → Run Maintain

Prototype	Proof of Concept	Pilot	Enterprise	Run Maintain
Functional Governance	Light Governance	Architecture Review	Enterprise Governance	

Activities (Prototype)

- Assessing opportunities
- Identifying value owner
- Determining proof points to test

Activities (Proof of Concept)

- Verifying understanding of the problem
- Testing possible solutions
- Inspiring end-users to build confidence

Activities (Pilot)

- Making business decision using the tool
- Building robust, reliable end-to-end solution
- Verifying value of data and algorithms

Activities (Enterprise)

- Scaling-out technology solution
- Embedding business change & support
- Applying life-cycle management

Activities (Run Maintain)

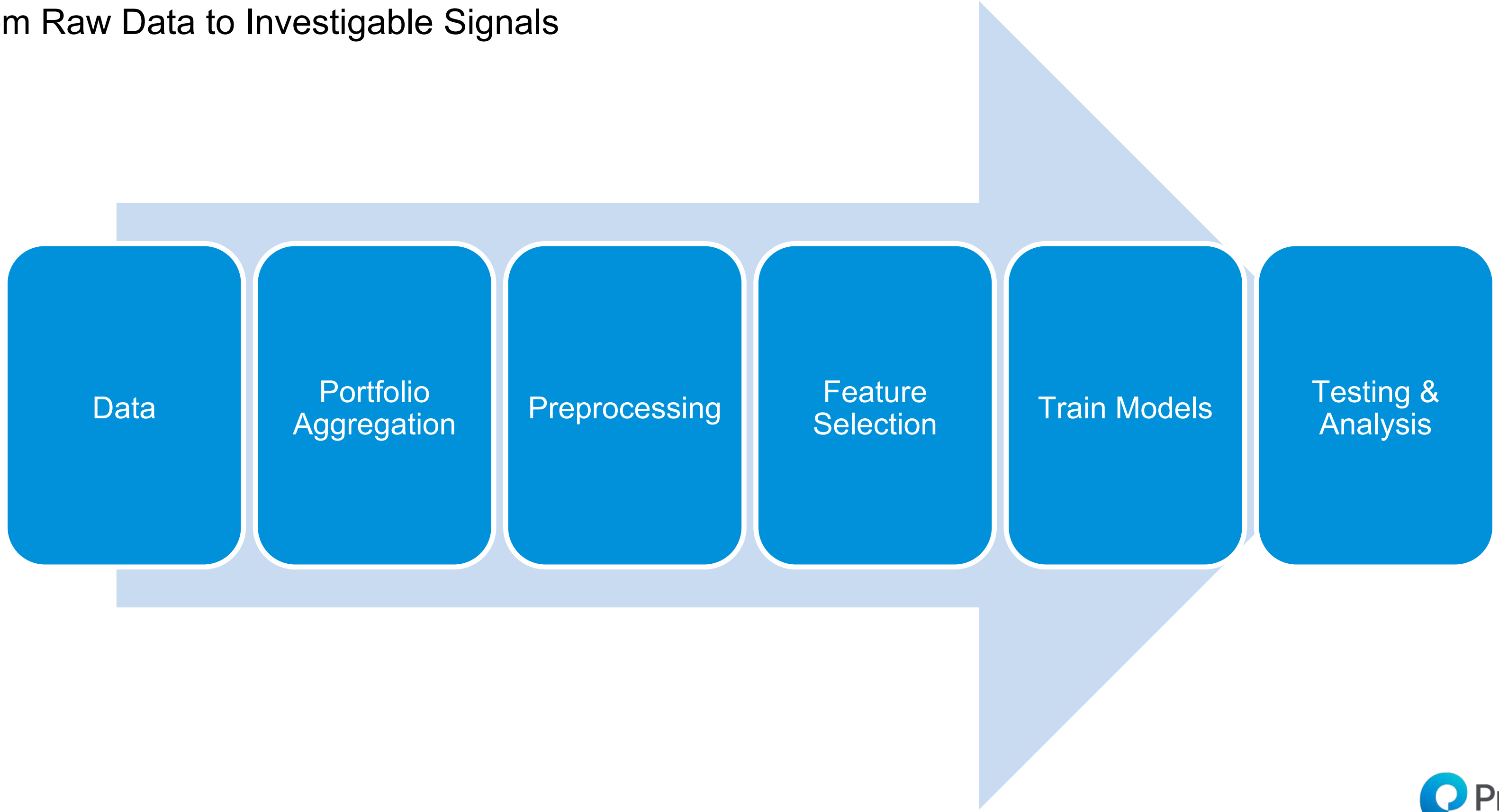
- Run & Maintenance of application landscape
- Bug fixes and minor enhancements
- Operational performance
- Disaster recovery

Market Survey

- Predicting the direction of stock market prices using Random Forest
    - Paper published in 2016
    - In-depth discussion of the mechanics of Random Forest
    - Displayed very high accuracy for short term classification results
- Prediction Algorithms and Confidence Measures Based on Algorithmic Randomness Theory
    - Paper published in 2002
    - Introduction to confidence measures in classification models
    - Achieved about 99% accuracy in classifying handwritten digits using SVM with confidence measures.
- Principal Component Analysis
    - Paper created in 2017
    - Demonstrates the usefulness and potential for PCA
    - Shows Cross Validation techniques to improve models

Stock Market Prediction Flow

From Raw Data to Investigable Signals



Prediction Pipeline

Data Aggregation

- *Stock-level to Feature-level*
- *Split data into sorted deciles*
- *Aggregate decile feature data using mean, median, standard deviation and variance*
- *Create prediction features (e.g. Vol, CSV) from algorithmic definition*

Feature Analysis

- *Reduce total number of features for model input*
- *Create new data from feature reduction algorithms*
- *Select features that have the most effect and less dependence*
- *Learn more about how each feature contributes to the model*

Machine Learning Models

- *Use aggregated data as training/testing data*
- *Experiment with regression machine learning models*
- *Continue improving models that work well*
- *Replace poor performing or ill-suited models with new experimental models*

Portfolio Predictions

- *Take latest market data and provide predictions for future performance*
- *Use predictions to maintain a test portfolio*
- *Organize results into a working pipeline that encapsulates the stages of our process*

Agenda

- Project Introduction
- Project Scope
- **Results**
- Looking Forward

Prediction Pipeline

Motivation

- What
  - Python implementation of modeling process
  - Specifications for each step
  - Provides framework for automating modeling process
- Why
  - Facilitate continuation of our exploratory research
  - Organize process into well-defined modules
  - Provide prototype of modeling automation

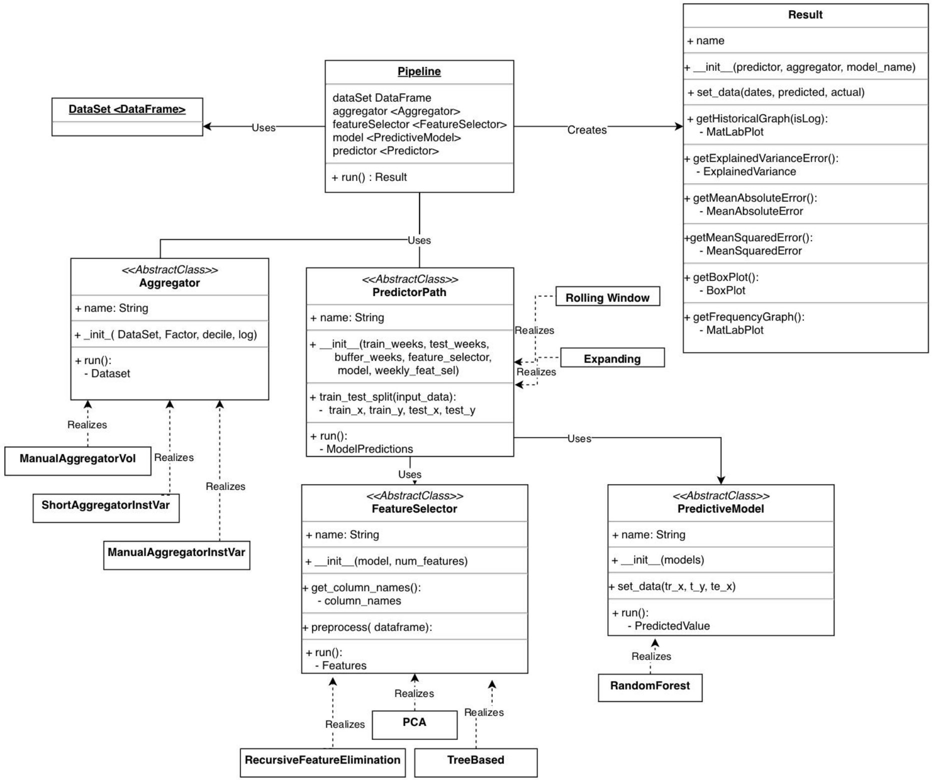
Prediction Pipeline

Requirements

- Functional Requirements:
  - Must run each step of the prediction pipeline in order without additional user input between steps
  - Must provide a detailed result that includes details of the pipeline execution and results of the model-fitting.
- Non-Functional Requirements:
  - Components must be general enough that new components may be created and used easily
  - Must be able to provide results in a reasonably short length of time
  - Must include documentation for easier extensibility
- Constraints:
  - Implemented in Python and/or R
  - Use NumPy and Pandas for implementation

Prediction Pipeline

Class Diagram



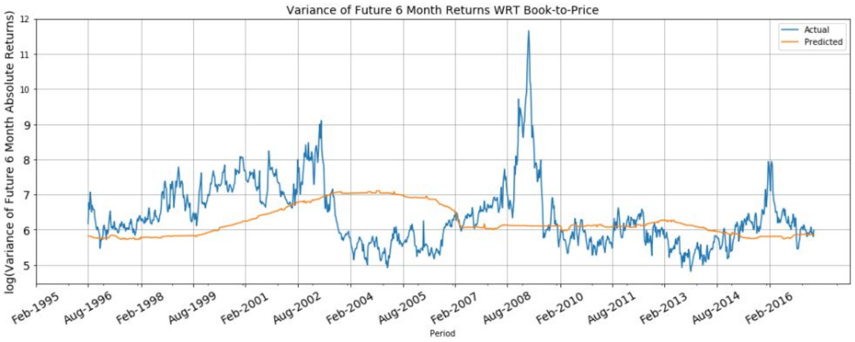
Prediction Pipeline

Design and Implementation

- Library Components:
    - Pipeline
    - Pipeline Component
    - Aggregator
    - Feature Selector
    - Predictive Model
    - Prediction Path
    - Result

```
aggregator = ManualAggregatorInstVar(merged_data, fac, ret)

feature_selector = RecursiveFeatureElimination(RandomForestRegressor(), 50)

predictive_model = IsolationForestModel(IsolationForest(n_estimators=50))

prediction_path = RollingWindowPath(3*52,
                                    52,
                                    26,
                                    feature_selector,
                                    predictive_model)

pipeline = Pipeline(m= predictive_model,
                    a= aggregator,
                    f= feature_selector,
                    p= prediction_path)

result = pipeline.run()
```

Prediction Pipeline
Historical Analysis



Gradient boosting model with loss calculation of least absolute deviation with learning rate 0.001.

Random forest model with number of estimators set to 100.

Challenges

Problem	Mitigation Strategy
Initial Infrastructure Issues *Memory issues for concurrent users*	• *Use swap space as supplementary memory to increase maximum synchronous usage*
Domain Research vs. Working Towards Deliverables *Lacking domain knowledge, but making good progress and results*	• *Deliverables have taken precedence over research* • *Find subject matter experts to assist our learning*
Accurate Feature Selection and Elimination *Needed to reduce the amount of features to get realistic and interpretable results*	• *Feature selection phase integrated into workflow*
Ensure Models are Developed Correctly *Ensure that we miss as little as possible to make a realistic and accurate model*	• *Ask the Principal team questions on the dataset and output from models* • *Look at output statistics other than accuracy*

Principal®

Models & Algorithms

Feature Analysis/Selection

Feature Analysis Method	Effect
Recursive Feature Elimination	*Performs a greedy search to find the best performing feature subset. Iteratively creates models and determines the best or the worst performing feature at each iteration.*
Principal Component Analysis	*Creation of new axes based on eigenvalues to explain the most amount of variance with the least amount of features. Primarily a dimensionality reduction technique.*
Tree-based Feature Selection	*Tree-based estimators are used to compute feature importances, which in turn are used to discard irrelevant features*

Models & Algorithms

Machine Learning

Predictive Model	How it Works	Advantages
Random Forest Regression	Creates a “forest” of decision trees Evaluates a decision based on splits between trees Analyzes results from several decisions to produce prediction	Classification or Regression Simple and flexible Easier to understand Quick to develop working models Quick to see out-of-box scores
Support Vector Machine	Supervised learning classification technique which aims to create decision boundaries by maximizing distance between a hyperplane and each of the classes.	Model non-linear decision boundaries Effective in high-dimensional datasets
Auto-Regressive	Use previous outcomes in a time series to predict future outcomes	Useful for cyclic data or highly autocorrelated data
Gradient Boosted Trees	Takes in a regressor and builds an additive model. This model is then tested against a loss function. The regressor is then fit on the output and direction of the loss function to optimize the learning of the regressor	Utilizes an extra factor to help it learn with optimization of the model

Agenda

- Project Introduction
- Project Scope
- Results
- **Looking Forward**

Project Conclusion

Project Highlights

- Learned about and successfully applied machine learning models within the financial domain
- Created a prototype forecasting application (Pipeline) with high extensibility using Python
- Provided new insights and useful results to our client that will provide value to the company

Project Future

- All code will be transferred over to Principal Financial Group for future usage
- May be developed further into a fully fleshed out application for portfolio analysts to use regularly
- Our work will lay the groundwork for future stock analysis techniques within the organization

Thank You!

Questions?

Principal®