**1数据获取**

数据获取可以分为 API数据获取、网络爬虫获取和文件获取三种方式，每种方式都可以开发出独立的模块。

**2.**[**数据预处理**](https://www.zhihu.com/search?q=%E6%95%B0%E6%8D%AE%E9%A2%84%E5%A4%84%E7%90%86&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7B%22sourceType%22%3A%22answer%22%2C%22sourceId%22%3A2974240137%7D)

数据预处理包括从收集的数据中提取有用的特征、处理缺失值、数据规范化等。pandas 库在这方面非常有用，可以帮助快速处理数据。

[**特征工程**](https://www.zhihu.com/search?q=%E7%89%B9%E5%BE%81%E5%B7%A5%E7%A8%8B&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7B%22sourceType%22%3A%22answer%22%2C%22sourceId%22%3A2974240137%7D)

特征工程是指从原始特征中提取新的特征，并使用这些特征来创建一个更好的[预测模型](https://www.zhihu.com/search?q=%E9%A2%84%E6%B5%8B%E6%A8%A1%E5%9E%8B&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7B%22sourceType%22%3A%22answer%22%2C%22sourceId%22%3A2974240137%7D)。一些常见的技术包括移动平均、[指数平滑](https://www.zhihu.com/search?q=%E6%8C%87%E6%95%B0%E5%B9%B3%E6%BB%91&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7B%22sourceType%22%3A%22answer%22%2C%22sourceId%22%3A2974240137%7D)、[技术指标](https://www.zhihu.com/search?q=%E6%8A%80%E6%9C%AF%E6%8C%87%E6%A0%87&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7B%22sourceType%22%3A%22answer%22%2C%22sourceId%22%3A2974240137%7D)（如 RSI）等。可以使用 ta 库来计算这些技术指标。

**划分[数据集](https://www.zhihu.com/search?q=%E6%95%B0%E6%8D%AE%E9%9B%86&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7B%22sourceType%22%3A%22answer%22%2C%22sourceId%22%3A2974240137%7D" \t "_blank)**

需要将数据集划分为[训练集](https://www.zhihu.com/search?q=%E8%AE%AD%E7%BB%83%E9%9B%86&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7B%22sourceType%22%3A%22answer%22%2C%22sourceId%22%3A2974240137%7D" \t "_blank)和[测试集](https://www.zhihu.com/search?q=%E6%B5%8B%E8%AF%95%E9%9B%86&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7B%22sourceType%22%3A%22answer%22%2C%22sourceId%22%3A2974240137%7D)，以便于[评估模型](https://www.zhihu.com/search?q=%E8%AF%84%E4%BC%B0%E6%A8%A1%E5%9E%8B&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7B%22sourceType%22%3A%22answer%22%2C%22sourceId%22%3A2974240137%7D)的效果。可以使用 sklearn.model\_selection 中的 train\_test\_split [函数](https://www.zhihu.com/search?q=%E5%87%BD%E6%95%B0&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7B%22sourceType%22%3A%22answer%22%2C%22sourceId%22%3A2974240137%7D)进行划分。

** 选择模型**根据问题的性质，选择合适的机器学习或深度学习模型。可尝试使用线性回归、[支持向量机](https://www.zhihu.com/search?q=%E6%94%AF%E6%8C%81%E5%90%91%E9%87%8F%E6%9C%BA&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7B%22sourceType%22%3A%22answer%22%2C%22sourceId%22%3A2974240137%7D)（SVM）、[随机森林](https://www.zhihu.com/search?q=%E9%9A%8F%E6%9C%BA%E6%A3%AE%E6%9E%97&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7B%22sourceType%22%3A%22answer%22%2C%22sourceId%22%3A2974240137%7D" \t "_blank)、XGBoost、长[短时记忆](https://www.zhihu.com/search?q=%E7%9F%AD%E6%97%B6%E8%AE%B0%E5%BF%86&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7B%22sourceType%22%3A%22answer%22%2C%22sourceId%22%3A2974240137%7D" \t "_blank)神经网络（LSTM）等模型。可以使用 [sklearn](https://www.zhihu.com/search?q=sklearn&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7B%22sourceType%22%3A%22answer%22%2C%22sourceId%22%3A2974240137%7D) 或 tensorflow.keras 等库来创建所需的模型。

** 训练模型**将处理好的数据输入到模型中进行训练。需要设置合适的[超参数](https://www.zhihu.com/search?q=%E8%B6%85%E5%8F%82%E6%95%B0&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7B%22sourceType%22%3A%22answer%22%2C%22sourceId%22%3A2974240137%7D" \t "_blank)，如学习率、迭代次数等。使用 fit 方法进行训练。

** 评估模型**在测试集上评估模型的性能。可以用[均方误差](https://www.zhihu.com/search?q=%E5%9D%87%E6%96%B9%E8%AF%AF%E5%B7%AE&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7B%22sourceType%22%3A%22answer%22%2C%22sourceId%22%3A2974240137%7D" \t "_blank)（MSE）、均方根误差（RMSE）、[平均绝对误差](https://www.zhihu.com/search?q=%E5%B9%B3%E5%9D%87%E7%BB%9D%E5%AF%B9%E8%AF%AF%E5%B7%AE&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7B%22sourceType%22%3A%22answer%22%2C%22sourceId%22%3A2974240137%7D" \t "_blank)（MAE）等指标来评估模型的效果。可以使用 sklearn 中的 mean\_squared\_error、[mean\_absolute\_error](https://www.zhihu.com/search?q=mean_absolute_error&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7B%22sourceType%22%3A%22answer%22%2C%22sourceId%22%3A2974240137%7D" \t "_blank) 等函数计算这些指标。

** 预测**使用经过训练的模型进行股票价格预测。可以使用 predict 方法得到预测结果。

** 优化**根据评估结果对模型进行[优化](https://www.zhihu.com/search?q=%E4%BC%98%E5%8C%96&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7B%22sourceType%22%3A%22answer%22%2C%22sourceId%22%3A2974240137%7D" \t "_blank)，如调整超参数、增加更多的特征、使用不同的模型等。

**** [**可视化**](https://www.zhihu.com/search?q=%E5%8F%AF%E8%A7%86%E5%8C%96&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7B%22sourceType%22%3A%22answer%22%2C%22sourceId%22%3A2974240137%7D)使用 matplotlib 或 plotly 等库对股票价格和预测结果进行可视化，以便于更直观地观察模型的效果。

**2.数据研究**

数据研究可分为以下六种方式:

(1)根据股价和成交量来分析价格的变化趋势。

(2)根据股票基本数据、产业行业数据、财政政策等来分析选择股票或预测股票的价格趋势。

(3)股票价格和成交量指标的分析，主要通过数学公式变换和指标图形研究股票价格的变化规律。

(4)理论研究，通过对周期理论、波浪理论、分形理论、周易阴阳理论等的研究预测股票的价格趋势。

(5)量化分析，即利用因子模型、多维分析来预测股票的价格趋势。

(6)基于深度学习来预测股票的价格趋势。

**3.策略研究**

通常，仅仅通过进行技术分析还不能保证资金池的收益最大化，因此还需要结合策略研究。如果把证券交易看作战争，那么数据研究就属于战术层面的研究，而策略研究就属于战略层面的研究。策略研究的重点是使风险最小化，同时使收益最大化。战略研究的目的是能赢，而不是赢多少。

**4.历史回测**

历史回测是使用历史行情数据进行回测分析来检验操作策略是否有效，以便对策略参数进行调整。

**5.策略执行**

由于策略执行涉及自动交易问题，因此其基本上为纯计算机技术问题。策略执行一般通过两种方式来实现:API的接口和外挂方式。