# 模型的原理

**ARIMA**

ARIMA是指自回归综合移动平均模型。它是一种用于分析和预测时间序列数据的流行统计方法。ARIMA模型是建立在三个主要组成部分上： 自回归（AR），整合（I），和移动平均（MA）。

该模型的AR部分是一个回归模型，根据一个时间序列的过去值预测其未来值。它假定一个时间序列的当前值是其过去值的线性函数。自回归的程度由AR模型的阶数（p）决定。

该模型的I部分用于通过对数据进行差分来消除时间序列数据的趋势。这样做是为了使数据静止，这意味着其统计属性（如平均值和方差）不随时间变化。差分的程度由积分模型的阶数（d）决定。

模型的MA部分是一个回归模型，根据时间序列过去的误差（残差）来预测时间序列的未来值。移动平均的程度由MA模型的阶数（q）决定。

参数p、d和q是通过分析时间序列数据的自相关和部分自相关函数确定的。一旦参数被确定，ARIMA模型就被用来对时间序列的未来值进行预测。

总的来说，ARIMA模型是时间序列预测的有力工具，特别是对于具有静止或准静止结构的数据。

**LSTM**

长短期记忆（LSTM）模型是一种递归神经网络（RNN），旨在处理传统RNN中的梯度消失问题。它被广泛用于时间序列预测，因为它能够捕捉到长期的依赖关系，并且能够抵抗过度拟合。

LSTM模型的基本原理是保持一个记忆单元，可以在很长一段时间内存储和检索信息。这是通过一组调节信息流入和流出记忆单元的门来完成的。三种类型的门是：

遗忘门： 这个门决定记忆单元中哪些信息应该被遗忘或丢弃。它从先前的隐藏状态和当前的输入中获取输入，并为存储单元中的每个元素输出一个介于0和1之间的数字。0的值意味着 "忘记这个信息"，1的值意味着 "记住这个信息"。

输入门： 这个门决定哪些新信息应该被储存在记忆单元中。它从上一个隐藏状态和当前输入中获取输入，并为记忆单元中的每个元素输出一个介于0和1之间的数字。数值为0意味着 "不存储此信息"，数值为1意味着 "存储此信息"。

输出门： 这个门决定记忆单元中的哪些信息应该作为当前的预测结果被输出。它接受上一个隐藏状态的输入和当前的输入，并为存储单元中的每个元素输出一个0到1之间的数字。数值为0意味着 "不输出此信息"，数值为1意味着 "输出此信息"。

这些门由可学习的参数控制，这些参数在训练过程中被更新。LSTM模型可以使用时间反向传播（BPTT）进行训练，以最小化预测值和实际值之间的预测误差。

总的来说，LSTM模型能够捕捉到长期的依赖关系，并通过保持一个记忆单元和使用门来调节信息流，对时间序列数据进行准确的预测。

**XGBoost**

XGBoost是一种集合学习方法，利用决策树来模拟输入特征和目标变量之间的非线性关系。它被设计用来处理各种类型的数据，包括时间序列数据。

XGBoost算法迭代地建立了一个决策树序列，每次都将一棵新的树与前一棵树的剩余误差相匹配。该模型使用梯度提升框架，这意味着它从一个简单的模型开始，并通过添加更多的树来逐步改善它。在每次迭代中，该算法计算损失函数相对于前一个模型输出的梯度，并将一棵新树与损失函数的负梯度相匹配。

在时间序列预测中，XGBoost可以通过在历史数据上训练模型来预测一个时间序列的未来值。输入特征可以包括目标变量的滞后值和其他相关变量。XGBoost还可以处理缺失数据，并能自动处理特征选择。

总的来说，XGBoost是一种强大的机器学习算法，可用于时间序列的预测，并在许多基准数据集上显示出最先进的性能。

**贝叶斯优化算法提升模型效果的原理**

通过对模型的目标函数进行不断地采样，构建一个代理模型来估计目标函数，然后利用代理模型进行优化，从而找到最优的超参数。

# PPT的细节

**PPT的里的图很难看看出你的模型好不好**

这里我解释一下预测图很难看出结果主要是因为项目选取的模型在时间序列预测任务里面都是合理有效的，所以很难看出效果的差别。但是可以发现细节的地方（指着图）比方说有很多时间点预测值离标签值是比较远的，仔细留意这些不拟合的地方就可以比较出这些模型的好坏。另外我们也提到了衡量模型好坏的指标体系。

**为什么组合模型准确率在结果中为什么比不过XGBoost？（单一XGBoost的缺点）**

事实上MAE不是唯一的指标，我们的指标体系里面组合模型的绝对误差、相对误差和平均相对误差都是最小的。结果中组合模型的准确性可以与XGBoost相比，甚至在某些情况下超过了XGBoost。在我们测试的所有模型中，我们提出的组合模型取得了最好的预测性能，这一点综合所有指标可以看出。

然而，需要注意的是，模型的性能可能取决于各种因素，如数据的质量和数量、输入变量的选择、模型参数以及评估指标的选择。因此，我们需要仔细评估该模型在不同情况下的表现，并做出适当的调整以提高其准确性。

总的来说，我们提出的模型对小麦期货价格预测领域是一个有价值的贡献，因为它可以提供更准确的预测，可以帮助农民和决策者做出更明智的决定，减轻小麦价格波动带来的风险。

**单一ARIMA模型进行时间序列预测存在以下缺点：**

**对非平稳时间序列的适应性差：**ARIMA模型要求时间序列数据是平稳的，如果时间序列数据不平稳，则需要进行差分处理，将其转化为平稳时间序列。但是，对于某些非平稳时间序列，即使进行差分处理，仍然难以转化为平稳时间序列，这时ARIMA模型的预测效果可能会受到影响。

**对异常值敏感：**ARIMA模型对异常值比较敏感，而时间序列数据中常常存在异常值。如果异常值未被正确处理，可能会导致模型预测结果出现较大偏差。

**单一LSTM模型进行时间序列预测存在以下缺点：**

**对于长期依赖关系处理存在困难**：虽然LSTM模型能够捕捉到时间序列数据的长期依赖关系，但是当时间序列数据存在较长的时间依赖时，单一LSTM模型的预测效果可能会受到影响，因为它可能会面临梯度消失和梯度爆炸的问题。

**对于缺失数据的处理有限**：LSTM模型在训练时需要连续的时间序列数据，如果时间序列数据中存在缺失值，需要对其进行插值处理或者删除缺失值。但是，插值可能会引入噪声，而删除缺失值可能会导致数据量减少，影响模型的准确性。

**对于多个时间序列的处理有限**：LSTM模型通常处理单个时间序列，对于多个时间序列的处理有限。例如，如果需要考虑多个产品的销售量，可能需要建立多个LSTM模型，每个模型对应一个产品，这可能会导致模型数量过多，难以管理。

**计算费时**。

**单一XGBoost模型进行时间序列预测存在以下缺点：**

**忽略时间相关性**：XGBoost是一种决策树算法，没有显式的考虑时间相关性。时间序列数据具有时间相关性，因此忽略了时间相关性可能会导致模型的精度下降。

**无法捕捉长期依赖**：时间序列数据中通常存在长期依赖，也就是当前时间点的观测值受前面很多时间点的观测值影响。XGBoost模型只能考虑当前时间点之前的一些观测值，无法捕捉到较远时间点的依赖关系。

**对于季节性变化的处理有限**：某些时间序列数据具有季节性变化，例如每年夏天销售额增加。XGBoost模型可以通过引入季节性特征来处理这种情况，但是只能处理已知周期的季节性变化，无法处理非固定周期的季节性变化。

因此，对于时间序列预测问题，单一XGBoost模型存在局限性，需要结合其他方法进行改进和优化。例如，可以考虑引入集成其他模型，以更好地捕捉时间相关性和长期依赖。也可以使用集成学习方法，例如将多个XGBoost模型进行组合，以提高模型的准确性和稳定性。

**请更清楚地举例子说你们工作的市场价值：**

农业产业： 对小麦价格的准确预测可以帮助农民、贸易商和其他行业参与者做出更明智的决定。例如，农民可以决定出售小麦作物的最佳时间，贸易商可以调整他们的交易策略以实现利润最大化，食品加工者可以更好地管理他们的投入成本。这可以导致农业产业内效率和利润的提高。

政府政策： 政府可以利用小麦价格预测来做出有关农业政策和干预措施的明智决定。例如，政府可以调整进出口政策，保持小麦价格稳定，确保公民的粮食安全。这可以帮助稳定小麦市场，缓解价格波动。

金融机构： 准确的小麦价格预测也可以使金融机构受益，如银行和投资公司。这些机构可以利用预测结果来管理与小麦相关的投资风险，如商品期货和期权合同。这可以帮助减少潜在损失，提高投资回报。

总之，模型的服务对象很广泛，可以渗进各行各业。行业参与者受益，包括农民、贸易商、食品加工厂、政府和金融机构。通过提供更准确的小麦价格预测，该模型可以帮助提高农业行业内外的效率、盈利能力和稳定性。

# 反驳性

**你们的项目目前在市场上有多少相关工作？你的优势在哪里？**

经过调查市面上的确有很多类似的工作，但是我认为我们的成果是独特的。不仅是把这三个模型结合起来进行预测，而且使用的指标体系和优化方法也是我们的独特之处。

我们的模型还有（1）易于利用；（2）处理时间快；（3）使用方便等优势。我们也后续也可以考虑将训练好的模型放到网络平台上去离线使用。

**我不认为这个工作有多创新？**

其实论文的创新都是站在巨人的肩膀上的，比方说给某个模型添加一个模块，修改训练过程之类的。那我们这种对多种模型的整合本身就是一种创新。一种优于任何单独模型的混合模型，可以突破每个单一模型的局限性，来创造更多的社会价值。这种整合方法还比较新，在时间序列预测中没有被广泛采用。其次就是我们在应用上本身就是创新，在铺天盖地的机器学习工作中，太少用在农业上面了。粮食属于民生大计，针对这一点进行研究我认为不仅具有创新性而且意义重大。

**你的组合模型会不会太过时了？**

目前不算太过时。在我们选题之时，这种组合模型的效果比我们找到的大多数时间序列预测模型都要准确。但是AI领域日新月异，如果不定期更新和改进，模型会很快变得过时，不断评估和完善模型以确保它们在结果准确高效是至关重要的。所以而在这之后我们也会不断追随潮流，用更新更高效的模块和方法去改进我们的模型的。

**数据有哪些特征**

在PPT的第8页有提到，可以读一下。

**XGBoost怎样整合前两个ARIMA,LSTM的预测结果？**

可以使用堆叠方法，将ARIMA和LSTM模型的预测结果作为新特征，将其与其他特征一起输入到XGBoost模型中进行训练和预测。具体而言，可以使用ARIMA和LSTM模型的预测结果作为输入，然后通过交叉验证等方法训练一个XGBoost模型，该模型将ARIMA和LSTM模型的预测结果作为新特征，从而可以更好地利用它们的信息。在预测时，可以分别使用ARIMA和LSTM模型预测得到的结果，然后将它们作为新特征输入到XGBoost模型中进行预测。

**GPT的摘要总结：**

该项目旨在建立更准确的小麦期货价格预测模型，以反映小麦市场的情况，保证小麦最低收购价政策的顺利实施。该项目利用中国国家粮油信息中心、《中国农产品价格调查年鉴》和《中国统计年鉴》的公开数据，从国际、国内两个市场探索影响小麦价格的重要因素。随后，项目根据随机森林模型对特征的重要性排序，构建了影响小麦期货价格的综合因素指标体系。项目在构建的指标体系基础上开发了传统的时间序列分析模型、单变量LSTM模型和单变量XGBoost模型。此外，项目还开发了多变量LSTM和多变量XGBoost模型，构建了更准确的小麦期货价格预测模型。该项目通过使用堆叠集合学习算法，结合三个模型的预测值，使用贝叶斯优化算法训练SVR模型，改进了现有的预测模型。项目利用绝对误差、相对误差、平均相对误差等各种指标对复合模型、单一模型、单变量模型和多变量模型的预测结果进行了比较，并选出了表现最佳的模型。最后，项目根据模型的有效预测结果提出了政策建议，这有助于及时调整市场供求关系，稳定粮食经济。