1:35

Alexey

**This week, we'll talk about stock market analysis with Python and machine learning. We have a special guest today, Ivan. Ivan works as Analytics Team Lead at Google. He's interested in investing in programming in Python, and he blogs at PythonInvest.com, so check out his blog. Welcome, Ivan.**

1:57

Ivan

Hi Alexey. I'm really happy to be here. I can give a short intro of my professional experience.

2:08

Alexey

**Yeah, we'll go into that. I just want to first thank Johanna Bayer, who prepared the questions for this interview. Also, before we go into our main topic of stock market analysis with Python and machine learning, we also need to add a disclaimer here. We usually don't do this, but since this topic is about stock market analysis, it's kind of a sensitive topic in the sense that, if you invest something, you do this at your own risk. We're not experts here – we're not certified experts in the field. We cannot claim that this is… I don't know, if…**

**There is a disclaimer that I probably need to read verbatim, so I'll just do that. It sounds boring, but just to be on the safe side. “Given that we are going to discuss financial matters, we want to add that we are not certified experts in the field. We do not claim that the material discussed in this podcast is complete, nor correct, nor a full disclosure of all risks. Neither should anything discussed in this podcast be considered a solicitation, recommendation, or offer to buy or sell any security financial product or instrument.” [chuckles] There is one more sentence. “Private advisors should be consulted before entering any transaction.” I had to say that. Thanks, Johanna, for actually thinking about that and preparing it. Anyway, this is a serious topic, and also quite interesting.**

# Ivan’s background

3:53

Alexey

**Before we actually go and discuss this topic, I want to know about your background. Can you tell us about your career journey? What did you do in your life?**

4:06

Ivan

I’ll probably start a little bit earlier. I finished universities – one was about computer science and another one was Master of Arts and Economics, specializing in finance and data analysis. Since then, I always worked between analytics, programming, and real business. I started my professional experience while working for four years as a financial analyst at a big commercial bank. There, I learned how to deal with SQL Server, how to write queries. I also built a few profitability models for banking products like loans, mortgages, etc.

Then I joined Google in Moscow and I worked there for two and a half years. The position was called Industry Analyst. I worked with a large sales team that covered only the Russian market, but the position was international straightaway, so there are many similar analysts in any country in the world and we shared our experience very often. I worked there for two and a half years and then I decided to move to Ireland. In Ireland, I also changed my position from an analyst to a consultant.

This consultancy job was focused on mobile web. I learned how to optimize UX and speed for mobile websites, how to pitch new technologies, like progressive web apps, accelerated mobile pages, and best user experience in filling lengthy forms. After two more years, I changed my position to my current role. I returned to the analytical role – I work for a large sales team, again, covering a slightly different product. But now the role is about all European markets, Middle East, and Africa. It's even more global. We do participate in global product managers meetings and planning. That's, that's my short professional experience.

6:42

Alexey

**So you mostly worked at Google.**

6:47

Ivan

For almost ten years now – nine and a half years.

6:50

Alexey

**Ten years – that's really impressive.**

6:53

Ivan

I feel like a dinosaur here. [chuckles] Many, many new people come and leave, but I'm still here.

7:02

Alexey

**It's amazing if you can find a place where you can constantly grow. It's impressive – growing without feeling bored.**

7:14

Ivan

Yeah, it's a lot of opportunities and changing positions – three positions at Google – helps as well, because it's all different roles.

7:24

Alexey

**And now, as an Analytical Team Lead, you probably also manage people. You work as a manager.**

7:29

Ivan

Not really. It was an experiment for a half year. I prefer to be an IC (individual contributor). For me, it is easier to focus on complex projects, while being a people manager is a very different role. I try to stay away. I was offered it a few times, but I want to drive IC projects for now.

7:56

Alexey

**You didn't like it?**

7:59

Ivan

It is very different. I think I lose focus [in this role]. If you need to talk with 10 people, and it's weekly catch ups – you just can't drive any individual projects. I have so many things that I want to do, like complicated ML projects, or just analysis. So I prefer to drive direct business impact.

8:31

Alexey

**So, as an analyst, you can do machine learning projects?**

8:36

Ivan

I can. It's not very often for a business intelligence analyst and they are hard to do. But there are no restrictions. If you really want to do this, you can do it.

8:53

Alexey

**Okay, that's cool. I see why you stayed there for so long. Because there's such a variety and you don't really have any limits. If you want to do machine learning, nobody will tell you, “Hey, you're an analyst. You're not supposed to do that.” You just go and do it, right?**

9:07

Ivan

If you can prove that there is a business impact – if you can simplify and make your stakeholders interested, and this is actually something that is very valued by direct clients, and you bring incremental revenue, they will be happy to have this.

# How Ivan became interested in investing

9:25

Alexey

**How did you become interested in investing? Was it related to your education in economics or to something else?**

9:34

Ivan

Yeah, I'd say it started while I had a few courses on banking, financial markets, asset price – I read a lot of research papers about it. I think it was a good and a bad thing. A bad thing, because we had an idea that all markets are efficient and you can't get any arbitrage, so I didn't use any of that knowledge for some time. But I continued reading some books…

10:08

Alexey

**May I stop you? What you said, “All markets are efficient and you cannot use any arbitrage.” [Ivan agrees] [chuckles] What does this mean? What is it? Arbitrage is buy cheaper, sell more expensive?**

10:27

Ivan

Kind of. Risk less revenue, so that if you want to trade, and if you want to make any gains – so if you do it without any risk and with 100% probability, that may be considered as an arbitrage. There are scientific papers that say that markets are weakly or strongly efficient – that all of the arbitrage is taken away by big hedge funds, big companies. As an individual investor, there is no or very small opportunity to do that. But still, there are many people who are trading for their personal investments. They can make some profits, like 5–7% yearly profits in S&P 500, in the last 20 years, which is probably much higher than you can get from just having your money in the bank – 0–1%.

11:35

Alexey

**Okay. So you were interested in that, but you thought that markets are efficient, there is no arbitrage, so you didn't do any investing. But then you figured out that it's actually possible, right?**

11:47

Ivan

Yeah. At some point, I wanted to have a pet project. I wanted to test new technologies and wanted to have something on the internet, like create a website, publish some articles, record videos. My natural selection for this pet project was to create a blog about investing. After writing articles for a year or so, I started doing real investing.

12:23

Alexey

**Ah, so you weren’t investing. First, you were learning – you were blogging about that – and then you understood, “Okay, now I know enough. Let's actually try that.”**

12:33

Ivan

I had some theoretical knowledge from the university, but then I wanted to check all of that. I wrote programs in Python, managed to get the data, and I could see within simulations that I can have some profits – even expected profit. So I said to myself, “Now I should do this. I should try it.” Actually, this works as a self-enforcement and self-commitment, so that if I invest my own money, I will spend time on this, even if I don't have time left or no power left from my normal workday.

# Getting financial data to run simulations

13:15

Alexey

**Can you tell us more about this? You said you figured out how to get data and how to do simulations. Maybe you can tell us more about where you got that data from? And maybe where people can get data now (today) and how to perform these simulations?**

13:36

Ivan

Yeah, sure. I can say that my early article about getting data [which was written about] three years ago is still the most popular article on the website. So please check it out on PythonInvest.com – I think the main data source…

13:54

Alexey

**What’s the name of the article?**

13:56

Ivan

It's something about… Let me check. One second.

14:08

Alexey

**Exploring Finance APIs with Python?**

14:10

Ivan

Yes. Yes, that's right.

14:14

Alexey

**I will add the link now to [the description].**

14:21

Ivan

The very first link, Yahoo Finance API, is the most important one. You can get dailies, or even hourly stats about many stocks – and not only stocks on the market. In most cases, this is enough to start. On top of that, you can add some other APIs, like Quandl API or Pandas Data Reader, or paid APIs like Polygon.io if you need this. But I can say that it's quite hard to combine all of the data sources into one dataset. Normally, I have fragmented ideas where I use just a few data sources, and I want to build some strategy on that.

15:14

Alexey

**So what is this Yahoo Finance API? For each stock, do they give you a time series or what exactly?**

15:23

Ivan

Yes, there is a website finance.yahoo.com and this API… I think this is not even an official one – someone has been managing that for a number of years. The data looks like a daily or hourly or minutes time series, and it is called, OHCLV, (Open, High, Close, Low, and Volume). These five numbers perfectly indicate any time period – it can be a minute or an hour or a day. Sometimes, if you're really professional, you will need to connect a real time stream and have trades in your data, but it's not easily available for a retail investor. And you probably don't need it, because it's a lot of data. It's quite hard to get real-time decisions on that.

16:27

Alexey

**And you said this is kinda an official data source. What does that mean?**

16:32

Ivan

I'm not sure that it’s supported by Yahoo Finance. Maybe they scrapped it, but you can check in pibi – there is a GitHub repo with 2000 stars. Someone who is not directly related to Yahoo Finances… [cross-talk]

16:53

Alexey

**The API is supported by Yahoo, but the Python package is not, right? The Python package talks to the API.**

17:02

Ivan

You're probably right, yes. I don't know for sure. What I saw was that sometimes it is not stable and it has limitations. You can’t do more than 2000 calls in one hour. So if you want to scale your algorithm, or if you want to have it updated automatically every hour or every minute, you will probably have to move to paid data sources at some point.

# Open, High, Low, Close, Volume

17:28

Alexey

**But this is enough to get started. [Ivan agrees] I just checked. It's a web page – there is day, open, high, low, close, volume. [Ivan agrees] I have no idea what these things mean: Open, High, Low, Close. How much background do I actually need to get started here – to make sense of this data?**

17:53

Ivan

I think now, it's like a commodity that you can download via an app – you can invest 50 euros, 100 euros, whatever. You can buy [stocks], and you will quickly understand. So Open, it's the opening price, when the trade day begins. Close is the closing price. High is the highest price during the day. Low is the lowest price. And volume is just the volume of transactions – it can be the volume of money that was traded. Within these five numbers, they won't reflect the whole time series – it can be millions of trades for one stock in a day. It’s just five numbers, but generally, they are quite good.

18:39

Alexey

**And there is this “adjusted close” – what does it mean?**

18:45

Ivan

I do not know, I think it can be…

18:48

Alexey

**It’s actually the same as close. I don't see any difference.**

18:52

Ivan

Yeah, if you see “adjusted” you always use adjusted. For our purposes, I think it’s when you do analysis and trade not one second before previous data received, it doesn't… It's not very important because some time will pass, like a few minutes, a few seconds, and the price will be different for sure. So whether it is adjusted or not, it's not that important.

19:21

Alexey

**Okay. Right now, I opened your article, and in this article, there is a section about Yahoo Finance API. There is a link to finance.yahoo.com and a quote from Pfizer Inc. I see all these numbers: Open, High, Low, Close Volume. [Ivan agrees] What kind of analysis can I do on this data? How can I use it?**

19:47

Ivan

For example, you're interested in some stock or in some vertical that you're following (a few stocks), and you can treat it as a time series that can be growing, neutral, or declining. You can try to make profits from any trend that you can see. The simplest strategy can be the mean reversion strategy, where if you see that your stock is going down, and at some point, you will expect it to be back again to the previous historical mean (which doesn’t always happen, but oftentimes, it happens) and you can build some simple (or not that simple) strategy on these time series. You want to predict when it goes back and you want to buy lower and sell higher. That's how you make some profits.

20:47

Alexey

**If I understand the strategy correctly, what I do is – I have this data (I have high and low) and I probably need some sort of average for the day, so maybe it's just the average between high and low. I have that for each day. Then I see what the overall mean of the time series is. Then I see, “Okay, right now it's going back.” Then I can expect, with certain probability, that it will go back to the mean. Right now it’s below the mean, it will go back to the mean, so now I can buy the stock and expect that it probably should go back.**

21:32

Ivan

Yes, that's correct.

21:35

Alexey

**While an individual stock might not go back, overall, if you apply this [strategy] to 10-20 stocks, then it will kind of average out and the overall effect will be that most of them go back to the mean, while some of them don't.**

21:51

Ivan

Yeah. That's why you need a simulation and that's why you need prediction. You can’t guarantee 100% that it will go back, but if you're correct in 60-70% of chances, and your winning deals are actually covering your losses, then you make some profits.

# Risk management strategy

22:14

Alexey

**I have two questions right now. First question is, “Okay, now I see losses. What do I do with them?”**

22:24

Ivan

That's a very good question. [chuckles] I started with ML trading and I didn't have any risk management system implemented. Risk management normally means that, if you see any losses, you need to have some tolerance for losses – it can be 5%, 15%, 50% losses (if you trade crypto). You shouldn't be trading with your gut feeling, but if there are any losses higher than your threshold, you probably need to sell your position straight away and fix the losses.

Again, it depends on your strategy. You need to simulate and test that. If you trade with this loss fixing strategy or not, it should show you that you have better returns overall. Because you may sell when it went just below a 10% drop and then it came back. Maybe sometimes it is better to sell, sometimes it is better to wait. There is no real answer, but what I can say is that many traders (professional market participants who treat this as a full-time job) have these risk systems integrated so that they don't need to think when and how to trade. It's all algorithmic.

23:54

Alexey

**What if I don't sell? What if I don't recover the losses? It means I will not be able to use the money from selling the stock to buy something new, right?**

24:04

Ivan

You will just receive less money. You probably will be able to sell it if markets are not closed. Sometimes, when there's a big volatility, markets are closed, and you can't do anything. But in 99% of cases, especially if it's big markets, like the American Stock Exchange, they're always open – you will just receive less money and it will be much harder to return to something that you had previously.

24:46

Alexey

**I've just heard somewhere (I think it was a podcast that I was listening to) where they were talking about investments and they said that the best strategy is to always hold on. Okay, you have some shares, you see that they're dropping, just don't sell them. They will come back in two years, maybe, and you will still earn money from them.**

25:13

Ivan

Yeah, that's generally a good recommendation if you're not a professional trader, or if you do not have daily trading bots. If you're a long-term investor, you shouldn't be looking for short fluctuations that can be irrational [or] that can be connected to some news published. So if you're doing these accidental behavioral moves, generally, they are bad. If you're just an average person, it's always better to have a long-term strategy.

That's what I did when I didn't have a risk management system and when I didn't sell some of the stocks that started to drop – I had to sell them at some point of time and I had losses – I changed my strategy into a passive strategy, where I have a few funds (large, exchange-traded funds) that I invest in and they do not rebalance them. I decided once what should be a portfolio allocation and now, every time I have free money that I want to invest, I just buy exactly the same weights for a small number of stocks – and they are long, very long, so I hold them.

# Testing your trading strategies

26:48

Alexey

**Okay, I actually got lost a bit. You said many things, like “long stocks,” “rebalancing strategy,” “portfolio allocation”… We should probably talk about that later. [Ivan agrees] One question I still have is – we know how to get data from, for example, Yahoo Finance. Then, we talked about doing some analytics on that, and the easiest analytics is this mean reversion strategy. [Ivan agrees] We can see that if something is below the historical mean, we can expect that it will recover.**

**Then, we also talked about risk management strategies like, for example, how much we can tolerate when the price is going down? We have all these components, which is probably enough to get started. Then there is this important thing you mentioned, which is a simulation. I have some pieces in place. I think, overall, it will work. I have a strategy, I have a risk management strategy. How do I actually test it? How do I verify that if I put real money into this thing, it will not deplete my bank account?**

28:11

Ivan

Yeah, that's a usual thing when you deal with train/test validation datasets. The only thing that you can't do as a data engineer or a data scientist, is that you can't have them sorted in a random order. I always have a validation dataset, which is the latest data that is seen. The model never sees this data. If you do a simulation, you need to make sure that the model didn't see not only those numbers, but numbers around it – one hour, one day before it and one day after – because it can quickly learn and it will be a data leakage.

28:58

Alexey

**Can you say it again? You mentioned that there is some model – some strategy. It can be a simple analytical strategy – it can be a model. We use the data, for example, from Yahoo Finance, to come up with an algorithm or train a model. What you said is that we should leave the last few days (the last couple of records) aside – we don't train the model on them. I guess this is the usual kind of time series validation strategy, right? [Ivan agrees] What do you do with these last two records? You test your model on them?**

29:44

Ivan

Yes. I can give an example of the exact thing that I had. I started from the 100 largest US stocks, and I made predictions for one week ahead. I tried to predict… Historically you can calculate future growth from the data – when you don't know the future growth, but you have 20 years of history of so. So I built an ML model, I predicted growth for every stock, I combined some of the predictions with some rules, like thresholds so that the predictions should be strong enough. Then I picked the top three and invested with equal shares, but I didn't use the last one-two years of data for training and testing.

When I had that model, I used that model to predict and simulate using the last two years of data to see whether this strategy actually delivers me profits. I saw some variation – there were both profits and losses – but I calculated the rolling financial profit and some loss, and I wanted to see this rolling in the last six months so that I can consistently earn some money. This can be like 5% or 10% of my capital so that even if in one day, I can be plus 20%, on another day, I can be minus 10% – overall, on average, I can consistently earn every single month. That's an ideal picture to have.

31:34

Alexey

**So I'm just trying to picture this in my head. Let's say we have these largest US stocks. We have data for them for ten years. Let's say we train our model using six years – we take the last six years, not all ten years. We train our model, and we predict what the performance for the next week would be on historical data. [Ivan agrees] Four years ago or whatever.**

**Then we retain our model, we maybe move one day, or one week, and we kind of slide this window throughout the entire history of the data that we have. Every time we see – for the next week, for the next week, for the next week, how the model would perform with all the historical data we get from Yahoo. Right?**

32:32

Ivan

You need to separate the data you're training on and everything that you're simulating on. In the real setting, if you have some model that you need to train on, it just never sees this data. You can simulate using historical data. If you try to apply that, you train the model [on data from] one week ago or one month ago, and you show it the newest data. That's why I just leave one or two years’ data away and I do not use it for model training or for hyperparameter tuning or for anything else. After I train the model, I use this model to predict for the last one or two years. I simulate that. I trade with this model within some strategy that I have in my head, and I can check if it gives me some profits, then I would expect that it will continue working now.

33:42

Alexey

**Let’s say that today is 2023 and you want to train a model and come up with some strategy. You take some data for the last 10 years and, let's say, you train only from 2013 to 2021. Right? [Ivan agrees] And then you apply your model for the rest of the years – for 2022 and 2023. Then, on a daily basis, you see how much the strategy would yield or lose? [Ivan agrees] Then you also see what the overall yield of this is – is it positive or negative? How much you would earn.**

34:28

Ivan

That's why you need to compare – it should be higher than any other alternative investments that you can get. If you can get 3% [return from] the bank, then there is no point to do any investment trading if you get less than that. So if you need to have 5% or 10% or whatever you can get. Another thing is – there can be 100 different strategies to trade on using exactly the same predictions. So you need to actually run several simulations, and you need to be very deep in the stock markets and improve those strategies. They are equally as important as building a good model.

35:15

Alexey

**How do we actually do the simulation? We also need to think… What is a strategy? A strategy is – we have some algorithm, and then also what exactly we do with the output of this algorithm. In your case, you said, “We take the 100 largest US stocks, and we pick the three most performing ones. Then we have some sum of money, and we allocate it evenly – we split the sum of money among these three stocks.” That's a strategy, right?**

35:46

Ivan

Yes. Actually, there are five trading days for stocks (Monday to Friday) and at a specific time. Late in the evening of a previous day, I got all the data on the closed markets, and I could predict on the latest available day, what the returns for the next week [would be]. On the very beginning of the next day, when the market just opened, I bought some shares that were predicted to grow. But before buying the shares, I actually sold shares that I bought one week ago.

So it was a split, where every day, out of the five days, I would trade 20% of my capital – say I had 1000 euros to invest, I would trade 200 euros every day on a number of the predictions that I have. And if during my previous week, my investments actually delivered some gain, it will be not 200, but it can be 210 or 190 that is available for me just today. That was the exact strategy. I actually made predictions every single day.

37:11

Alexey

**What is not clear is… At the beginning, you allocate only, let's say, $1,000? It's a part of the strategy how exactly you split this 1000, right? [Ivan agrees] At the beginning, you can maybe take all of this 1000 to invest in these three companies, and then see how exactly it performs after one week, sell something, and with the money you get from that, you buy something new. Right?**

37:41

Ivan

Yes, that can be the strategy if you invest only once in a week. If you just invest on Friday, and you wait until next Friday, that can be a good strategy. But if you want to trade every single day, and you have different predictions, and you hope that the more predictions you have, the more trades you have and the expected return is higher from more trades, then you want to trade more often. So if you trade every single day, then you need to invest not 1000, but 200, because you don't have more money.

# Sticking to your strategy

38:24

Alexey

**So let's say I'm getting started and I have my data scientist salary. I think “Okay, if I leave aside 50 euros every day, my strategy can be – for every day when the market is open, I see how I can invest these 50 euros (or $50). Also, in addition to that, the stocks somehow are performing, so I can maybe sell the ones that are doing quite well and re-invest the money.”**

38:58

Ivan

That’s not a very good thing, because you start to look at whether they are performing or not – if they are not performing, you're hoping that they will be back. That's why you need a risk management system so that when you're selling, even if it is a loss, you have money available to invest for the next predictions. Also, you're limiting the downside of your strategy. You can't lose something like 50% or 20% – you can only lose 10% that you've identified. But on the winning part – that's exactly the same.

If you see that the stock is growing in one month, in one week, and you're thinking “Oh, it will grow again. Why should I sell?” But your algorithm says that you need to sell in one week. The only exception (and it can be an adjustment to your strategy) is that, if exactly one week after, your algorithm predicts again, “This stock is selected out of 100 and it has a very high chance of growth.” Maybe you don't need to sell it and hold it for another week. But you need to have these predictions. By default, you're selling no matter what, in one week.

40:20

Alexey

**Okay. I see. So you're not holding – whatever you have, you sell, and you reinvest the money.**

40:26

Ivan

Yeah. That's the prediction. That's what people call “day trading,” or “short-term trading,” or “arbitrage trading,” whatever. If you have another strategy that is more like a passive investment that you allocate once and you hold for some number of years, that's a totally different area.

# Important metrics and remembering about trading fees

40:51

Alexey

**There is a question about metrics. Let's say we do this simulation. I guess for us, what we mostly care about is the money – return on investment. Right? [Ivan agrees] Whether it's positive or negative. If it's positive, how much, right? That's what we care about. Is there anything else that we should look at when doing the simulation or when we’re actually trading?**

41:17

Ivan

Yes. I think there are metrics – the usual metrics for ML scientists, like accuracy, precision, and AUC if it's a binary model. You can make a regression prediction and bets for not just a market movement up or down, but for exact revenue that you will get – 10% or 5%, or 20%. This is a much harder thing to do. But I would say that if you’re thinking only about binary models – and this is what is different in finance, that I care more about precision, rather than accuracy. I need to make my predictions right and only half of the predictions are very important for me – those that are predicted to grow.

That's why I target precision. Another thing that you need to consider is trading fees. Every time you trade, you will give away some money as fees for buying and selling. Actually, you need to minimize the number of trades, because those fees can eat up all of your revenue. That is a big consideration for a simulation piece of work.

42:49

Alexey

**So when simulating, we also need to take the price of the fees into account. So when we calculate the total return on investment, it includes the fees.**

43:03

Ivan

Yes, it can be as simple as that. If you trade 200 euros every day, fees can normally be 1% – it is two euros to buy, two euros to sell, so it is four euros total. You need to make at least four euros in profits in one week to cover your trading fees. So you just need to be not just positive, but positive higher than a specific threshold.

43:39

Alexey

**We talked about machine learning models, specifically binary classification models. What is typically the variable (the outcome) that we’re trying to predict? Is it whether the stock is going to grow?**

43:56

Ivan

Yes. I calculate the growth rate in one week – for the next week or for the next month, it depends on the time period that you're trading. Let's assume it's one week. I can see that if it is growing more than zero or more than five percent, then I define this outcome as a one (as positive growth). I have different definitions. If it's just growing up or down, there are very balanced classes – 50/50 more or less. It is easier to predict, but your financial result probably won't be that perfect because sometimes you will guess correctly on the stock movements, but your trading fees will eat all of your revenue.

That's why I try to predict growth that is higher than 5%, for example, in a week. Now it's an unbalanced problem – probably it's only 20% of cases that have this theme. It's harder to build a model to predict this and maybe even different features are important. But now I have, hopefully, a better financial result. But it all needs to be simulated.

# Important features

45:23

Alexey

**You mentioned features. I assume that everything we’re talking about is based on the data that we talked about, which is Open, High, Low, Close, and Volume, right?**

45:40

Ivan

Yes, this is the most fundamental dataset. It can be 100 stocks, it can be some other data sources, but this one is maybe 70% of all the features that are the most important ones.

45:55

Alexey

**How do you build…? Let's say, we want to build the simplest possible model for that, but still use machine learning – like logistic regression or something else. How exactly would we design the problem in order to predict this growth of 5% or more? What kind of features would we use? What kind of model would we use? How would we prepare the data?**

46:27

Ivan

Yeah. We get all the historical data available, we generate some features (whether the stock is growing 10 days in a row, whether there are some patterns like a huge draw up or anything else based on the historical time series) and then you treat every observation like only one observation that you have. Just for today, what information do I have based on today's data and all the previous information from a number of days, weeks, or years.

So I get like 200 features, whatever I can come up with, and I try to predict the growth of more than 5% in a week. I can generate this from the data because I see what happened in five days. But I do not show this to the model. It's something that I want to predict, “What will happen in the nearest future?” I treat these observational as independent, which is not correct – stocks are generally dependent on each other and also dependent day-to-day. It is a simplification, but it works somehow, and you can have some profitable strategies.

48:02

Alexey

**So with these features that we generate, we can train logistic regression or XGBoost or [cross-talk] ML models, right? We don't even need to use any specific time series models here.**

48:19

Ivan

You don't. I have plans to use LCTM or recurrent neural networks, but I really doubt that they are much better. You can tune simple neural networks with so many parameters (or XGBoost) so that you will probably have good baseline model quality, so that you can apply a number of strategies.

48:53

Alexey

**When it comes to time series, some time ago (I think it was a very long time ago) but I was quite actively following Kaggle competitions. It was before LCTMs really took off. Back then, the models that were winning in time series competitions were still XGBoost and so on.**

49:17

Ivan

There are still the same models on Kaggle. That was a good revelation for me – that you actually need to work a lot on your features and area understanding rather than a very complicated model that you can’t debug and that you can't understand what happens inside. That's why I spent a lot of time on feature explainability and trying to really feel those connections – why this or that feature is the most important one and what examples show that this feature actually influenced the stock market move?

49:58

Alexey

**So the features, from what I understood from our discussion, are simple statistics but calculated over a period of time. Let's say “Was the stock growing in the last 5 days, 10 days, 20 days, 30 days?” Right? Then you can calculate many different features like that, or it could be the percentage of stock growing, it could be an average stock price, it could be the trend – all these features – and then you just throw all these features into your model and you see what is important and what is not. Then if you figure out which features are useful, which are not.**

**You also said that explainability is important, so then you can analyze all the features yourself, like, “Hmm. It makes sense that the model goes for that feature because ‘reasons’.” You can understand, because you handcrafted this feature yourself, you understand how it's built, what it's doing, and you can relate to why XGBoost decided to go with this particular feature as the most important one. Right?**

51:09

Ivan

Yes. And that's why simpler models are better. Because for logistic regression, you have feature importance straightaway. For decision trees, you can see the steps that it takes. Especially if the model is wrong, you can see why it’s relying too much on this or that feature. You probably need to generate five additional features, and you have some bias, you are missing something that the model is not capturing (these big movements) and thus predicting incorrectly.

# Deployment

51:46

Alexey

**I’m also curious about deployment. We discussed all the strategies, we discussed simulation – for simulation, you don't need anything. But for the actual trading… Let's say, now we have a strategy, we tested it with simulations, it's showing good ROI – how do we actually start using this? Do I just simply have a cron job on my laptop, do I run an instance on AWS where I have this cron job? What are the possible options?**

52:16

Ivan

That's exactly why I enrolled in your course. [chuckles] It’s about all these deployments and cron jobs and lambda functions and running a web service. For me, it was just a trained model – I clicked the button every day and made some predictions and tried to think about it, read some news. I do not feel very confident in fully automated trading. I would like to have some manual steps of checking, while there are not many trades to be made. But ideally, yes, it should be a data flow – ETL or whatever – and cron job or Airflow to launch these cron jobs and recalculate, get all the new data that you have, make the predictions, use it for different strategies (or one strategy) that you have, gets the traits descriptions.

You can even have an API connection that can place trades for you. That is also very convenient because you can add additional things, like stop loss, that you can place straight away, like, “Buy me a stock for $100, but if it goes beyond 90, sell automatically.” When you place that trade, you don't think more about it. That's what I'm hoping to achieve in a few months from now.

53:54

Alexey

**Now, since you do this manually with the help of algorithms, you have to be really stick to your strategy, right? Because you see, “Okay, the model says I need to sell it.” Then you may start questioning it.**

54:13

Ivan

Exactly. That's not correct. Many, many books prove that you shouldn't be doing this if you participate in daily, weekly (regular) stock trading.

54:26

Alexey

**You just have to follow the recommendations consistently because this exact strategy is what you used for simulations. You know that the ROI of this strategy is that number, so you have to stick to that. You really have to be consistent.**

54:42

Ivan

You can select some period, like three months or six months, and if your strategy is not delivering that ROI anymore, you probably need to come up with another strategy. Or you may have different strategies – two, three, five strategies – and you can see which strategies work better in the current market conditions.

# How DataTalks.Club courses helped Ivan

55:05

Alexey

**You mentioned the course from DataTalks.Club – I assume it was the MLOps course? Which one did you mean?**

55:11

Ivan

I followed MLOps, but now you have ML Zoomcamp, which has a little bit of MLOps things inside as well. I used to have only binary and the logistic regression model. Now I have decision trees and XGBoost and neural networks – so now I have many more models and now I can say, “Okay, if I hypertune them, how important is this hypertuning? How does it influence not only accuracy, but also the final simulation results?” MLOps is an important thing and you have another course in January – I will sign up for it as well. [chuckles]

55:58

Alexey

**Yeah, that's the data engineering one. It's amazing to hear that you found this useful. It's really pleasant to hear. [chuckles] Since we're talking about the course – right now, the students of Machine Learning Zoomcamp are doing the project, and then there is another project that we will have in January. Maybe you have some recommendations. If somebody (some students) want to build the model using financial data, what's the simplest thing they can do? Just go to Yahoo Data, get CSVs, and go ahead and try that?**

56:36

Ivan

Yeah, sure. Or just install this app. Don't think too much – don’t read 100 different books. Try to design. If you're an analyst, you will see opportunities straightaway. If you follow one or two companies, or a specific vertical, you will think about it, but you need to do a simple step and start actual trading. Then you will have plenty of ideas of what needs to be done. All of the materials with the simplest but working ML models are actually a very good fit for this task. What is interesting with this problem is that you probably won't reach 80 or 90% accuracy ever – this is a real world problem. It is hard to solve and this is a challenge.

# Ivan’s site and course sign-up

57:29

Alexey

**In this episode, we probably have talked enough about different things that for students, just from listening to this episode, it's sufficient to go and build the first model, right? But if you want to learn more, would you say that your website, Python Invest is a good source. What can people actually find there on your website?**

57:53

Ivan

It actually follows the same path that I was exploring. I started with the financial APIs and then I thought, “Okay, news probably is important.” And I added a news API, but it was kind of messy, and it's very hard to trade on it, and it's sentiment analysis, which is not necessarily connected with real returns. Then I had earnings per share, which is a fundamental metric, which you can get through the free API's. I had to build the right scraper, which was actually from the Yahoo Finance website. Then I said, “Okay, I'll build some models. I need to allocate between different stocks, not equally, on the top three predictions.”

Maybe you can differentiate those three predictions – your model predicted that they will grow, but what is important is that sometimes you can limit your down movement just by investing 10% in one stock, 50% in another, and 40% in the third one, rather than investing 30/30/30. Then came portfolio locations. That's a natural flow and I have 13 different stories, and I tried to construct them with TLDR and with the major results first, so that you don't need to go very deep into the code and data sources, and you can understand the idea before you read the whole article.

59:29

Alexey

**I just included the link to our description and to the live chat. Ivan, if you have a few more minutes. [Ivan agrees] There's one more question I wanted to ask. I checked your website and I see that there is a thing called “course”. [Ivan agrees] It's a six week boot camp. Can you tell us more about that?**

59:52

Ivan

Yeah, I’ve had this link for a year or more and it's my hope that I can launch it soon. For now, it’s just an email subscription and I’m getting people interested. It's a constant trade-off for me, whether I should wrap up everything that I learned and package it into some course and spend a few months doing that, *or* should I go and write new articles and apply new models to my personal trading. What has been working so far for me over the last two years was personal trading.

For a course, I probably need some collaborators, or I just need to replicate your mechanism of peer-reviewed things. I really liked your course. Or I could collaborate with someone else (with some platform) to design that. So I want to think only about the content, but not about everything else. It turns out that you need to think about many things while doing that.

61:05

Alexey

**It’s a lot. [chuckles] There’s a lot to do.**

61:06

Ivan

Yeah. I heard that and I will probably wait until 500 people are subscribed, so that I can say, “Okay, next year from January, I will run it.” It’s not there yet, but I hope someday.

61:24

Alexey

**So if anyone wants to learn more about that, there is a course – sign up, and when there are 500 people, we will have a course.**

61:32

Ivan

It's already around 200.

61:36

Alexey

**Okay. So we just need 300 more. [chuckles] Okay. Thanks a lot. It's unfortunately time to wrap up for today. It was amazing. I learned many new things. Hopefully, everyone else also learned new things. Thanks for joining us today, for sharing. Remember, this was not financial advice. I'm just saying that in case somebody didn’t hear the disclaimer at the beginning. But I personally learned a lot. It actually looks like a good project for the students of the course. Thanks again. And thanks, everyone, for joining us today, too.**

62:15

Ivan

Thanks for having me. Thank you, Alexey.